Prediction of Aggregation Rate and Aggregation-Prone Segments in Polypeptide Sequences

Dissertation
zur Erlangung der naturwissenschaftlichen Doktorwürde
(Dr. sc. nat.)

Mathematisch-naturwissenschaftlichen Fakultät der
Universität Zürich

von
Gian Gaetano Tartaglia
aus
Italien

Promotionskomitee

Prof. Dr. Amedeo Caflisch
Prof. Dr. Andreas Plu̇ckthun

Zürich 2005

To my wife with Love

Contents

Contents **2**

1 Introduction **7**
1.1 Protein Molecules 7
1.2 A Computational Study on the β-Aggregation 12
1.3 Aggregation Rate Prediction 15

2 A Computational Study on the β-Aggregation **28**
2.1 Genetic algorithm optimization 28
2.2 Molecular dynamics 30
2.3 Lila's set up and performances 31
2.4 Fibril formation 35
2.5 β-Aggregation Matrices 35

3 The Role of Aromaticity, Exposed Surface, and Dipole Moment in Determining Protein Aggregation Rates
Protein Science (2004) 13, 1939-1941 **45**

4 Prediction of Aggregation Rate and Aggregation-prone segments in Polypeptide Sequences
Protein Science (2005) 14, 2723-2734 **49**

5 Organism Complexity Anticorrelates with Proteomic β- Aggregation Propensity
Protein Science (2005) 14, 2735-2740 **62**

6 Conclusions **69**

List of figures **72**

Summary

Nature has designed proteins to fold to specific three-dimensional structures and to function as structural chemical building blocks in living systems. The sequence of amino acids encodes the native structure as well the energy landscape in which the protein searches out conformations. The functional native state of the protein most often corresponds to the thermodynamic free-energy minimum conformation at physiological conditions. The discovery of protein aggregation diseases, however, where multiple proteins sacrifice contacts in the globular state in favor of inter-chain contacts with neighboring proteins, suggests that some proteins have aggregated states that are thermodynamically equal if not more favorable than the native state. Many of these aggregates have a common morphology known as an amyloid fibril. Amyloid fibrils are composed of an intramolecular β-sheet core running perpendicular to the fibril for some micrometers in length and a few nanometers in diameter. There at least 16 distinct human diseases that are associated with amyloid fibril formation. Although early attention focused on the possible toxicity of the amyloid fibrils, it is now accepted that the prefibrillar intermediates are the main toxic species in aggregates. It has also been found that under a variety of solution conditions, most proteins can form amyloid fibrils indistinguishable from those of the amyloid diseases. Because of the difficulties in obtaining detailed structural information by X-ray crystallography or solution phase NMR spectroscopy, computational approaches are needed to identify physicochemical properties of sequences involved in the β-aggregation.

In Chapter 2, aggregation properties of small peptides are investigated by using a genetic algorithm optimization in sequence space and molecular dynamics sampling of conformation space. As target structures for the optimizations we used the parallel and the antiparallel β-sheet conformations of three heptapeptides: 1632 different sequences were sampled, for a total amount of 81 μs of simulation. We found that in sequences selected for parallel aggregation the number of aliphatic and aromatic residues increases

almost monotonically, while the number of charged and polar residues decreases. The opposite is observed in sequences selected for antiparallel aggregation. Results of the genetic algorithm optimization represent an essential element in the derivation of the equations presented in Chapter 3 and 4.

In Chapter 3, we report a formula to predict the change of aggregation and disaggregation rate upon amino acid mutation. In the derivation of the model, we used simple physicochemical properties of amino acids, such as the polar and apolar water-accessible surface areas, the π-stacking interaction of aromatic residues, the dipole moment, the β-propensity, and the charge. To have the most general model possible, we did not use any parameters that need to be estimated from experiments. Although the application of the equation shows high correlation with experimental data, the model require the *a priori* knowledge of wild-type aggregation rates.

In Chapter 4, we report an equation that does need any information apart from the amino acid sequence and two environmental factors (i.e., temperature and concentration) to compute the aggregation rate. Our model was used to predict the aggregation rates of human muscle acylphosphatase, islet amyloid, α-synuclein, tau, glucagon, calcitonin, fibronectin, titin, and toll with a correlation of 95%. Moreover, the equation allows the calculation of the amyloid spectrum of a protein, identifying those segments which are involved in the β-aggregation. In addition, the model distinguishes between the parallel and antiparallel β-sheet organization within the fibrils and shows that mammalian and non-mammalian prion proteins have different amyloid-spectra.

In Chapter 5, we analyze complete proteomes of several eukaryotes to identify changes of β-aggregation propensity through organisms of different complexity. From *P. tetraurelia* to *H. sapiens*, we found that proteomes of multicellular and more long-lived eukaryotes contain fewer sequences with high β-aggregation propensity and more proteins with low β-aggregation propensity. We also observed that compared to random proteomes, natural proteomes are enriched in proteins with low β-aggregation potential as well as proteins with high β-aggregation potential. Such polarization indicates the dual evolutive requirement of intrinsically disordered proteins with low β-aggregation propensity as well as proteins with a stable fold which comes at the cost of higher β-aggregation propensity.

Zusammenfassung

Proteine sind komplexe organische Verbindungen von hohem Molekulargewicht, die an den unterschiedlichsten und fundamentalsten Prozessen lebender Organismen beteiligt sind. Um ihre Funktion auszuüben, müssen sich Proteine in eine eindeutige dreidimensionale Struktur falten, den sogenannten nativen Zustand, welcher einzig und allein durch die Aminosäuresequenz eines Proteins bestimmt ist. Der native Zustand ist die Übersetzung der genetischen Information und entspricht der Bedeutung der Sequenz in der Sprache der Proteine. Aus Sicht der statistischen Mechanik entspricht die Faltungsreaktion einer Diffusion eines Ensembles von Polypeptidketten auf einer trichterähnlichen Energielandschaft. In diesem Zusammenhang stellt die vollständige Beschreibung der freien Energielandschaft eines Proteins eine effektive Art dar, die thermodynamischen und kinetischen Aspekte der Proteinfaltungsreaktion zu beschreiben und öffnet den Weg zum Knacken des "Proteincodes".
Zur Zeit ist es noch nicht gründlich verstanden, welche Vorgünge zur Bildung von geordneten Peptid- und Proteinaggregaten führen.
Der Schwerpunkt dieser Dissertation war die Erforschung der physikalisch-chemischen Eigenschaften der Aminosäuren, welche die Beta-Faltblatt Aggregation begünstigen. Darüber hinaus wurden zwei parameterfreie Formeln für die Vorhersage der Aggregationsraten vorgeschlagen (Kapiteln 3, 4, und 5). Zu den Verfahren, die für die Herleitung der Modelle benutzt wurden, zählen unter anderem: die Analyse von beta-aggregierenden

Peptidsequenzen, welche durch im Sequenzraum optimierte genetische Algorithmen modelliert wurden und die Erforschung des Konformationsraums durch Moleküldynamiksimulationen (Kapitel 2).
Die beobachtete hohe Korrelation zwischen vorhergesagten und gemessenen Aggregationsraten weisen darauf hin, dass unsere Modelle in vitro Experimente mit hoher Genauigkeit beschreiben.

Chapter 1

Introduction

1.1 Protein Molecules

Amino Acids

Amino acids consist of a primary amine bound to an aliphatic atom (called the α-carbon, or C_α), which in turn is bound to a carboxylic acid group. The C_α bears a side chain which is different for different amino acids. Proteins are linear polymers of amino acids linked via a peptide bond which consists of a carbonyl group's carbon atom directly bound to the nitrogen atom of a secondary amide. The peptide chain has an unbound amino group free at one end (called the N-terminus) and a single free carboxylated group at the other end (called the C-terminus). While there are theoretically billions of possible amino acids, natural proteins are formed from only 20 amino acids (called proteogenic). The side-chains of proteogenic amino acids are quite varied: they range from a single hydrogen atom (as for glycine, the simplest amino acid) to bicyclic groups, as for tryptophan. In fact, the 20 amino acid side chains show different physicochemical properties such as polarity, acidity, basicity, aromaticity, bulkiness, conformational flexibility, ability to cross-link, ability to hydrogen bond, and chemical reactivity. These characteristics, many of which are interrelated, are largely responsible the wide range of protein properties.

The information contained in the amino acid sequence (called primary structure) is enough to guide a protein to fold into its three-dimensional structure (called the "native state") [1], to determine its specificity for interaction with other molecules [2, 3] and to set its lifetime and stability with

respect to the unfolded state [4]. The protein function is almost completely dependent on protein structure [5, 6]. Enzymes must recognize and react with their substrates with precise positioning of critical chemical groups in the three-dimensional space. Scaffold proteins must be able to dock other proteins or components precisely and position them in space in the correct way. Structural proteins like collagen must face mechanical stresses and be able to build a regular matrix where cells can adhere and proliferate. Motor proteins must reversibly convert chemical energy into movement in a precise fashion.

Protein Folding

Almost a half century ago, Pauling discovered two quite simple, regular arrangements of amino acids, the α-helix and the β-sheet, which have become known as the main components of the secondary structure of proteins [7]. In the protein core, the secondary structure provides a way to preserve hydrogen bonding of the peptide backbone by forming regular and repeating structures. The folding of secondary structural elements together with the spatial arrangement of the side chains is known as the tertiary structure. As many proteins are composed of two or more polypeptide chains which associate through non-convalent interactions and disulfide bonds, the description of the spatial arrangements of these chains is called the quaternary structure.

Folding depends a great deal on the characteristics of a protein's surronding solution, including the identity of the primary solvent (either water or lipid inside cells), the salt concentration, the temperature, and molecular chaperones [8]. In an aqueous environment, proteins fold in order to put as much of the hydrophobic amino acid side chains out of contact with water as possible. This provides much of the driving force for protein folding and protein-protein interactions. Generally, polar amino acid side chains partecipate in hydrogen bonding to water, while hydrophobic side chains interfere with it. In fact, protein structure (and also the interactions between proteins and other molecules) may be regarded as a compromise. On the one hand, it may be necessary to sacrifice a hydrogen bond or two to gain two or three hydrophobic interactions. On the other hand, it may be necessary to place a hydrophobic residue in contact with water in order to pick up a few more hydrogen bonds in the secondary structure [9, 10, 11]. Among the interactions that provide protein's stabilization [12, 13, 14] we can include:

- Hydrophobic interactions
- van der Waals interactions
- London dispersion forces
- Hydrogen bonds

- Charge-charge interactions

Among the unfavorable interactions we can highlight:

- Removing a polar group from water without forming a new hydrogen bond to it
- Removing a charged group from water without putting an opposite charge nearby or putting two like charges close together
- Leaving a hydrophobic residue in contact with water
- Putting two atoms in the same place (steric exclusion)
- Organizing anything into a structure (decreasing entropy)

For a 100-residue protein, it is possible to estimate roughly that the sum of all the favorable interactions that stabilize the three-dimensional, native structure is of the order of -500 kcal/mol . On the other hand, the sum of all the unfavorable interactions that destabilize the structure is approximately +480 kcal/mol. The net result is that the three-dimensional structure of a typical protein is only about -10 to -15 kcal/mol more stable than the structureless state. In fact, proteins can lose their structure if put in unsuitable chemical (e.g., high or low pH; high salt concentrations; hydrophobic environment) or physical (e.g., high temperature, high pressure) conditions [15, 16]. This process is called denaturation. Denatured proteins have no defined secondary and tertiary structures and, expecially if concentrated, tend to aggregate into insoluble masses. Many of these aggregates have a common morphology known as amyloid fibrils which are regular fibrillar structures that are micrometers in length, a few nanometers in diameter [17, 18]. Amyloid fibrils are composed of an intramolecular β-sheet core running perpendicular to the fibril axis [19, 20].

Experimental work in the mechanism of protein folding has been greatly influenced by Levinthal's famous paper of 1969 [21], in which he pointed out that a polypeptide chain would require an astronomical amount of time to explore at random all possible conformations in order to finally reach the native state [22, 23, 24]. This motivated the search for partially folded intermediates that guide the protein to the native state [25, 26]. Importantly, the slowest folding proteins require many minutes or hours to fold. However, small proteins, with lenghs of hundred or so amino acids, typically fold on a millisecond time scale [27]. The very fastest known protein-folding reactions are complete within a few microseconds [28, 29].

Thermodynamics and Kinetics

Since the late 1980s, a theoretical approach to protein folding has been the calculation of protein energy landscapes [30]. The energy landscape of a

protein is the variation of its free energy G as a function of its conformation, owing to the interactions between the amino acids residues. The free energy change ΔG is a balance of two terms:

$$\Delta G = \Delta H - T\Delta S$$

where:

- ΔH = enthalpy, i.e., net amount of energy available from changes in the bonding of the reactants and products. If heat is given off, the reaction is favorable ($\Delta H < 0$).
- ΔS = entropy, i.e., change in the amount of order during the reaction. Order is unfavorable ($\Delta S < 0$). Disorder is favorable ($\Delta S > 0$).

Protein folding and aggregation can be thought as chemical reactions, in which the evolution of the atoms over time provides a complete description of both the thermodynamics and kinetics [31]. Before a reaction can happen, the individual molecules must have enough thermal energy to make or break the chemical bonds as required in the selected chemical reaction. In fact, the reaction can be viewed as being blocked by a barrier: If the barrier is low, the reaction is fast but if the barrier is high, the reaction is slow. For a generic reaction $A \rightarrow B$ (examples include: A could be the native state of a protein and B its unfolded state, A could be the unfolded state of a protein and B its β-sheet fold, A could be the aggregated state of a certain amount of peptides and B their dissociated state, etc.) the energy that a protein must gain to cross over the barrier is called the free energy of activation. Transition state theory tells us that when a molecule of a reactant has enough energy to jump the barrier, the molecule's structure is intermediate between that of the substrate and that of the product [32, 33, 34]. Importantly, the activation energy determines how fast a given reaction happens. The speed, or rate, of the formation of B or the disapperance of A is usually found to be proportional to the concentration of A that is present at the time the velocity is measured:

$$\kappa = -\frac{d[A]}{dt} = \nu[A]$$

This equation is known as a rate law and relates how the rate of the reaction depends on the concentration(s) of the substrate. In this case k must have units of molar per second (M/s) and $[A]$ has molar units (M), therefore ν must have units of reciprocal seconds ($1/s$). Since the reaction rate decreases as the substrate is used up, the plot of $[A]$ against time is a curved line; the slope decreases exponentially with time:

$$A = A_0 \; e^{-\nu t}$$

where A is the concentration of the substrate A at any time t, A_0 is the initial concentration of A at $t = 0$, and ν is the rate constant.

Energy Landscapes

It has been proposed that natural proteins have evolved such that the complicated free-energy surface has a funnelled shape which leads towards the lowest free-energy conformation available to the protein [35]. The funnel landscape allows the protein to fold to the native state through any of a large number of pathways and intermediates, rather that being restricted to a single mechanism [36, 37, 30]. The theory is supported by computational simulations of model proteins and has been used to improve methods for protein structure and design [38, 39, 40, 22]. Experimental techinques provide much of the information relative to the free-energy landscape:

- Fluorescence and infrared sprectoscopy (IR) capture the early events in protein folding on a submillisecond time scale [41, 42].
- X-ray and NMR spectroscopy determine interactions between individual atoms providing insights on the location of active sites, catalytic mechanisms, and conformational changes. [43, 44, 45].
- Far-and near-UV circular dichroism (CD) determine the average content of the secondary structure and monitor the packing of aromatic side-chains that determine conformational changes in the protein during folding [46].
- Single molecule detection (SMD), and in particular atomic force microscopy (AFM), total internal reflection fluorescence microscopy (TIRFM), optical-trapping nanometry, polarized fluorescence and fluorescence resonance energy transfer (FRET) monitor the time evolution of single biomolecules during their functional activity allowing the detection of global movements and conformational changes [47, 48, 49].

Molecular Dynamics

Current experimental strategies are not sufficient to provide all the information required to describe the protein free-energy landscapes: X-ray and NMR experiments do not allow the complete exploration of non-native structures, CD and AFM techniques are generally limited by the time required or on the space resolution. One of the principal tools to determine protein structures from X-ray crystallography and from NMR experiments is the method of molecular dynamics simulations. This computational method calculates the time dependent behavior of a molecular system and provides detailed information on the fluctuations and conformational changes of proteins and nucleic acids. The core of the molecular dynamics algorithm is the the potential energy function (force field). Current generation force fields provide a reasonably good compromise between accuracy and computational efficency

and are calibrated to experimental results and quantum mechanical calculations of small model compounds. The development of parameter sets for the force field is a very laborious task that requires extensive optimization. Force fields show certain limitations. As an example, no drastic changes in electronic structure are allowed, i.e., no events like bond making or breaking can be modeled. Nevertheless, molecular dynamics has been successfully applied to study the reversible folding of structured peptides [50, 51, 52, 53].

In agreement with experiments *in vitro*, three replicas of the Sup35 yeast prion fragment `GNNQQNY` have been shown to form a β-aggregated structure by molecular dynamics simulations [54, 55]. The investigation of the free energy of the system has indicated a highly rugged surface with minima corresponding to β-aggregate structures. Moreover, the parallel configuration, in which all of the peptides have the same orientation, characterizes the global minimum. Other configurations corresponding to different peptide orientations (including the antiparallel arrangement) identify several free-energy local minima [54].

1.2 A Computational Study on the β-Aggregation

The above mentioned system of three β-aggregated heptapeptides is the object of the study presented in **Chapter 2**, in which molecular dynamics simulations are combined with a genetic algorithm optimization to investigate the β-aggregation propensity of several amino acid sequences. The work aims to determine the information encoded in the β-aggregated structures and thus identify the polypeptide sequences that are susceptible to having a stable minimum in the β-aggregated state. Intriguingly, all the sequence mutations that lead to the disaggregation of the system represent straightforward candidates for the inhibition of amyloidogenic natural proteins.

Broadly speaking, working with genetic algorithms has the potential to be a philosophically and epistemologically-interesting iterative process. In the first step, the process of evolution occurring spontaneously in nature is observed. Next, principles of evolution are converted into computer programs. To complete the recursive cycle, computational genetic algorithms are applied to the very objects, i.e., proteins, from which they were derived in the beginning. In fact, the genetic algorithm is a heuristic method that operates on pieces of information just as nature does on genes in the course of evolution. Individuals are represented by a linear string of letters of an alphabet (in nature they are nucleotides, in genetic algorithm they are bits, characters, strings, numbers or other data structures) and are allowed to mutate, crossover and reproduce. All the individuals of one generation are evaluated by a fitness function. Depending on the generation replacement

mode a subset of parents and offspring enters the next reproduction cycle. After a number of iterations, the populations consists of individuals that are well adapted in terms of the fitness function.

In our case, the genetic algorithm searches the space of sequences for the ones that have the best match to a particular three-dimensional target conformation (a parallel or an antiparallel β-sheet aggregate of three heptapeptides [54]). For each peptide sequence, three replicas are submitted to a 330 K molecular dynamics simulation, starting from the β-aggregated conformation. A non physiological temperature of 330 K is used to obtain enough sampling within the time scale of the simulations [54]. Peptide sequences are ranked according to their ability to prevent disaggregation. The disaggregation fitness function is estimated for each sequence to be the number of snapshots whose C_α root mean square deviation (RMSD) from the template is lower than 1 Å. The best matches, called best parents, are replicated and subjected to mutation and crossover: 1632 sequences have been studied for a total of 81 μs of simulation.

Amyloid Fibrils

The above mentioned system of three β-aggregated heptapeptides can be only distantly compared to the long, twisted, and intertwined amyolid fibrils found *in vivo*. As shown by experiments *in vitro* that use circular dichroism and Fourier transform intra-red spectroscopy (FTIR), amyloid fibrils have a high content of β-structure, even when the monomeric peptide or protein is substantially disordered or rich in α-helical structure [56]. Investigations of the fibrils using electron and atomic force microscopy show that they are typically straight and unbranched. The fibrils are typically 6–12 nm in diameter and usually consist of two to six protofilaments, each of a diameter $\sim$ $2nm$, which are often twisted around each other to form supercoiled rope-like structures. Each protofilament in such structures appears to have a highly ordered inner core, which X-ray fiber diffraction data suggest consists of some or all of the polypeptide chain arranged in a charactersitic cross-β structure. In this structural arrangement, the β-strands run perpendicular to the protofilament axis, resulting in a series of β-sheets that propagate along the direction of the fibril (Figure 1.1) [18, 57, 58] .

Amyloid Diseases

Incorrectly folded proteins are responsible for prion related illness such as Creutzefeldt-Jakob disease and Bovine spongiform encephalopathy (Mad Cow disease), and amyloid related illnesses such as Alzheimer's disease. These diseases are caused by misfolded proteins aggregating into insoluble proteins. The discovery of protein aggregation diseases, where multiple proteins sacrifice contacts with the native state in favor of inter-chain con-

tacts with neighboring proteins, suggests that some proteins have aggregated states that are thermodynamically equal if not more favorable than the native state. Until about 30 years ago, proteolysis was considered to be the primary factor for the formation of amyloid aggregates *in vivo*, following the demonstration that lysosomal enzymes, at acidic pH values, are able to convert amyloidogenic proteins into amyloid fibrils [61]. The perspective changed around 10 years ago when it was shown that thransthyretin can be converted *in vitro* into amyloid fibrils following an acid-induced conformational change [62]. In 1998, a protein unrelated to any amyloid disease was found to form structures indistinguishable from the amyloid fibrils produced from the disease-associated proteins [63, 64]. In fact, under certain conditions it has been shown that any polypeptide chain can form fibrils [65, 66]. The various peptides and proteins associated with amyloid diseases have no obvious similarites in size, amino acid composition, sequence or structure. Nevertheless, the amyloid fibrils into which they convert have marked similarities both in their external morphology and in their internal structure.

Prions: The most interesting example of a protein folding disorder is Mad Cow disease and its human equivalent, the Creutzefeldt-Jakob disease. These diseases, along with the sheep version known as scrapie, have had the scientific community in uproar for years. They are all infectious diseases which are transmitted by prions, i.e., protein particles. Prions seem to be pure protein; they contain neither DNA nor RNA. However, an infectious agent must by definition be self-replicating [67, 68, 69]. The protein whose aggregation damages cells in Mad Cow disease is constantly being produced by the body. Normally, it folds properly, remains soluble, and is disposed out without problem. When a small amount of it misfolds in a particular way so as to become a scrapie prion, this prion bumps into a normal-folding intermediate, shifts the folding progress in the scrapie direction and the protein, despite its perfectly normal amino acid sequence, also ends up as scrapie prion [70, 71, 72]. In this manner the process continues: So long as the body keeps producing the normal protein, a little bit of scrapie prion can keep on creating more and then even more of itself. In effect, the prion is replicating itself without needing any nucleic acid of its own. When seed quantities of two different scrapie prion strains are mixed in separate test tubes with large amounts of normal protein, each test tube produces more of the specific scrapie prion strain that was added. Each strain induces the normal protein to fold in exactly the same way as the original seed [73]. The strain breeds in the test tube, just as it does in the body.

Alzheimer's Disease: Alzheimer's disease afflicts 10 percent of those over 65 years old and perharps half of those over 85. In 1991, several different research groups found that individuals with specific mutations in their amy-

loid precursor protein developed Alzheimer's disease as early as age 40. The body processes amyloid precursor protein into a soluble peptide known as Aβ; under certain circumstances, Aβ then aggregates into long filaments that cannot be cleared by the body's usual scavenger mechanisms. These aggregates then form the β-amyloid, which make up the neuritic plaque in Alzheimer patients [74, 75, 76]. So the consistent association of amyloid preursor protein mutations with early-onset Alzheimer's disease has finally answered a long-debated question: the deposition of neuritic plaque is part of the pathway leading to the disease and not a late consequence of it. To help understand the Aβ aggregation process, researchers chemically synthesized fragments of the 40-amino-acid-long peptide [77, 78]. Specifically, the precursor fragments have to form a specific nucleus, which then propagates the amyloid process. Possibly the slowness of this first step is why Alzheimer's disease is almonst entirely limited to older people, and it could be that the mutations in amyloid precursor protein that lead to early-onset Alzheimer's are the ones that make it progress more quickly and easily *in vivo*.

1.3 Aggregation Rate Prediction

As explicitly mentioned for Alzheimer's disease, the speed of the β-aggregation represents a crucial factor in the evolution of all the amyloid-related pathologies [79, 80]. However, it has been recently shown that the amyloid fibril formation is not a property limited to a selected few proteins and that under certain conditions any polypeptide chain can form fibrils [65]. In fact, single amino acid substitutions have been used to investigate the fibril formation of the human muscle acylphosphatase protein (AcP) [66]: The small α/β protein was converted from a soluble and native form into insoluble amyloid fibrils in a solution containing moderate concentrations of trifluoroethanol. When analysed with electron microscopy, the AcP aggregate present in the sample after long incubation time consists of extented, unbrached filaments of 30–50 Å in width that assemble into higher-order structures. The fibrillar material was shown to possess extensive β-sheet structure as revealed by far-UV circular dichorism and infra-red spectroscopy. Furthermore, the AcP fibrils exhibit Congo red birefringence, increased fluorescence with thioflavin T and cause a red-shift of the Congo Red adsorbtion spectrum, which are characteristics typical of amyloid fibrils [81]. On the basis of this experiment, an empirical equation was proposed to explain changes of the aggregation rate upon amino acid mutations [82]. The high correlation found between rate and simple physicochemical properties (such as the β-propensity, charge, and hydrophobicity) is the original motivation for the work presented in **Chapter 3**.

The Relative Rate Equation

In **Chapter 3**, we report a fomula to predict the change of aggregation and disaggregation rate upon amino acid mutation. In the derivation of the model, we used simple physicochemical properties of amino acids, such as the polar and apolar water-accessible surface areas, the π-stacking interaction of aromatic residues, the dipole moment, the β-propensity, and the charge. The model is parameter-free, i.e., to have the most possible general equation, we did not use parameters that need to be estimated from experiments. Human muscle acylphospatase (AcP), islet amyloid polypeptide, prion peptides, α-synuclein, amyloid β-peptide, tau, leucine-rich repeat and other model peptides were used to test the equation: A correlation of 85% was found, indicating high agreement with experimental data (Figure 1.2). Furthermore, the equation was applied to predict the disaggregation rate of sequences generated by genetic algorithm optimization (see also Chapter 2) with a correlation of 80%. The fact that the model can be applied to describe both aggregation and disaggregation rates is a consequence of the very general functional form of the equation. Moreover, the absence of fitting parameters permitted the use of the same equation for the description of experiments *in vitro* and *in silico*, indicating that the method suites general application.

The Absolute Rate Equation

Although the application of the relative rate equation shows high correlation with experimental data, this models requires the *a priori* knowledge of the wild-type aggregation rate. In **Chapter 4**, we report an absolute rate equation derived from both first principles and analysis of aggregating sequences designed by a computational approach. The model gives both the aggregation rate and the 'amyloid spectrum' of a protein, identifying those segments involved in β-aggregation (Figure 1.3). Compared with models published by others [82, 83, 84], our equation is the only one which has been derived to predict β-aggregating segments in different polypeptide chains. In agreement with results obtained by genetic algorithm optimization in sequence space and molecular dynamics sampling of conformation space (see also Chapter 2), our model distinguishes between the parallel and antiparallel β-sheet organization within the fibrils, giving interesting insights into their structure. We also found that mammalian and non-mammalian prion proteins have different amyloid spectra. More specifically, the absence of the fragment SNQNN, present in mammalian prions, decreases the overall aggregation propensity of non-mammalian prions, indicating a species-specific behaviour consistent with experiments [85, 86] and supporting the hypothesis of a species barrier in the transmission of the prion disease [87].

Evolutionary Trends for the β-Aggregation

Intrinsically disordered proteins represent one of the major structural differentiations between proteins of mono- and multicellular eukaryotes [88, 89]. Unstructured proteins are largely present in higher eukaryotes, indicating that the native protein's function depends on a structural ensemble rather than an unique three-dimensional structure. Regions lacking specific three-dimensional structures have been associated with 28 distinguishable functions, ranging from DNA-binding to display of sites for phosphorylation to preventing interaction by means of excluded volume effects [90]. In **Chapter 5**, we report a novel approach to compare proteomes based on the statistical analysis of β-aggregation propensity. From *P. tetraurelia* to *H. sapiens*, we show that proteomes of multicellular and more long-lived eukaryotes contain fewer sequences with high β-aggregation propensity and are accumulated in protein with low β-aggregation propensity (Figure 1.4). We observed that compared to random proteomes, natural proteomes are enriched in proteins with low β-aggregation potential as well as proteins with high β-aggregation potential. Such polarization is a consequence of the dual evolutive requirement of intrinsically disordered proteins with low β-aggregation propensity as well as proteins with a stable fold which comes at the cost of higher β-aggregation propensity. The functional role of aggregation phenotypes in multicellular eukaryotes is still a matter of debate [91, 92]. In the future, we plan to use gene ontology annotations of proteins with high predicted β-aggregation propensity to obtain insights on the specific role of some of the amyloidogenic proteins of unknown function.

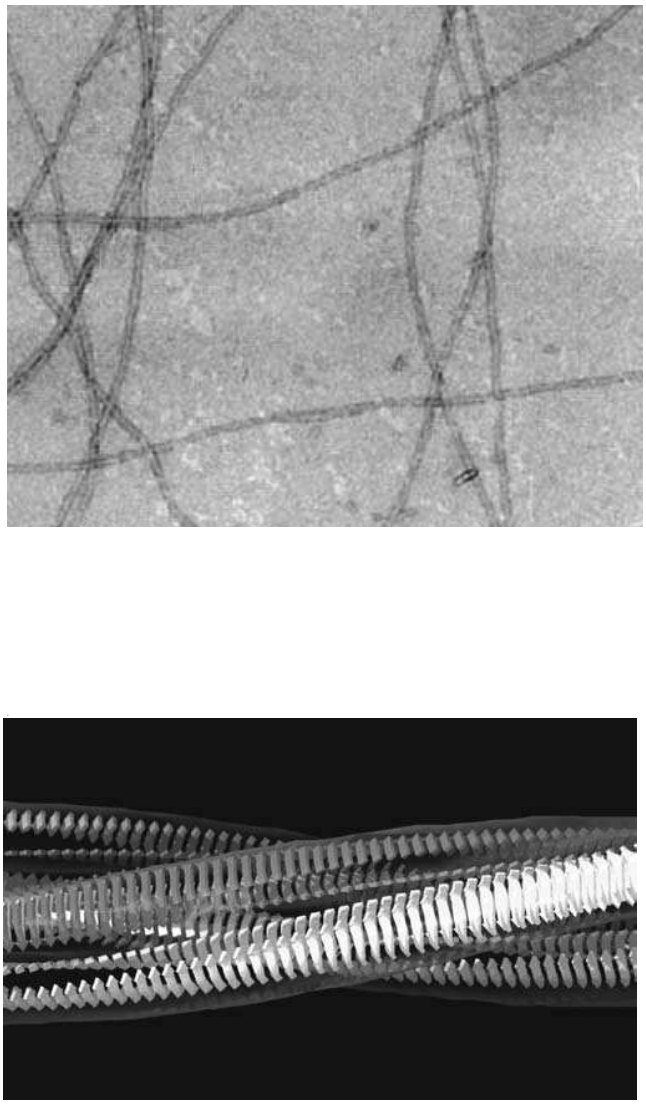

Fig. 1.1: *Top plot*: Transmission electron microscopy of a mesh of amyloid fibrils assembled from human lysozyme negatively stained with uranyl acetate [59]; *Bottom plot*: Schematic drawing of the structural organisation of insulin fibrils. The image shows a fibril with four protofilaments wound around each other. In this model the core structure of each protofilament is a row of β-sheets, here running antiparallel [60].

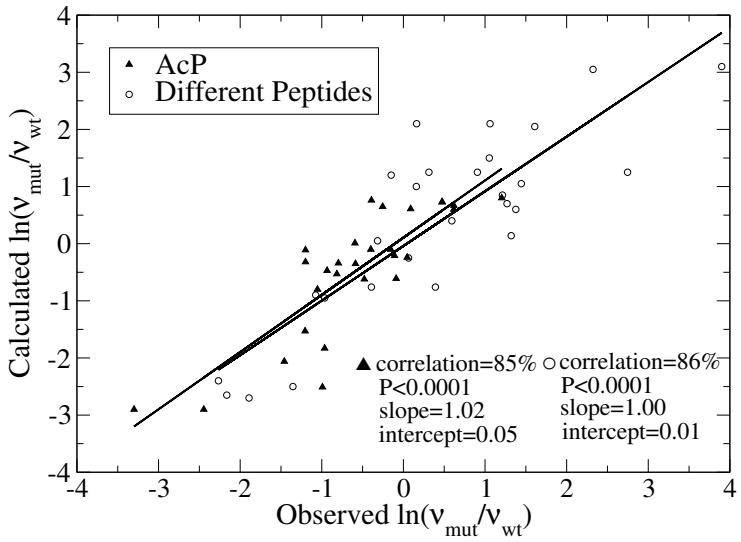

Fig. 1.2: Calculated versus observed changes in aggregation rate upon mutation: AcP (28 triangles) and heterogeneous groups of peptide and protein systems including islet amyloid polypeptide, prion peptides, α-synuclein, amyloid β-peptide, tau, leucine-rich repeat and some model peptides (27 circles).

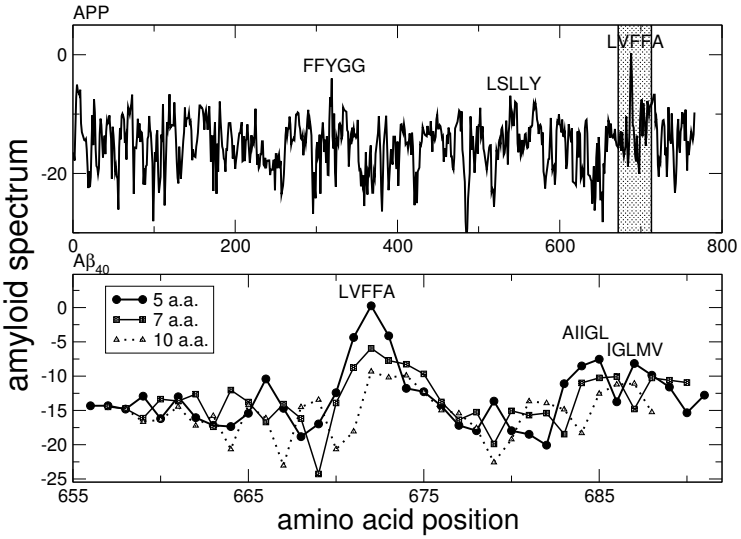

Fig. 1.3: Amyloid protein precursor. The amyloid spectrum is averaged over a window of five aminoacids. The entire sequence is scanned by shifting the window by one residue at a time starting from the N-terminus. The analysis shows a major peak corresponding to the segment `LVFFA` at position 671.

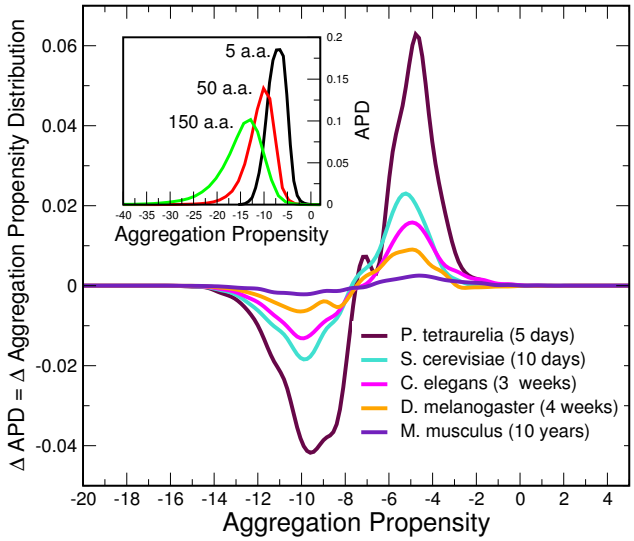

Fig. 1.4: *Inset:* Histogram of human polypeptide sequences as a function of β-aggregation propensity distribution at three different window sizes. *Main Plot:* Aggregation propensity distribution (APD) differences with respect to *H. sapiens* for complete proteomes of *M. musculus, D.melanogaster, C. elegans, S. cerevisiae* and *P. tetraurelia* (window size of 5 residues). Life-spans of organisms are reported in parentheses.

Bibliography

[1] Anfinsen, C. B. (1973) *Science 181*, 223–230.

[2] Koshland, D. E., J. (1958) *Proc. Natl. Acad. Sci. USA. 44*, 98–104.

[3] Janin, J. & Chothia, C. (1976) *J. Mol. Biol. 100*, 197–211.

[4] Warshel, A. & Levitt, M. (1976) *J. Mol. Biol. 106*, 421–437.

[5] Lacroix, E., Viguera, A. R. & Serrano, L. (1998) *Folding & Design 3*, 79–85.

[6] Fitch, W. M. & Margoliash, E. (1970) *Evol. Biol. 4*, 67.

[7] Pauling, L., Corey, R. B. & Branson, H. R. (1951) *Proc. Natl. Acad. Sci. USA. 37*, 205–211.

[8] Tomasi, J. & Persico, M. (1994) *Chem. Rev. 94*, 2027–2094.

[9] Hermann, R. B. (1972) *J. Phys. Chem. 76*, 2754–2759.

[10] Baldwin, R. L. (1986) *Proc. Natl. Acad. Sci. USA. 83*, 8069–8072.

[11] Jackson, R. M. & Sternberg, M. J. E. (1994) *Protein Engineering 7*, 371–383.

[12] Tanaka, S. & Scheraga, H. A. (1975) *Proc. Natl. Acad. Sci. USA 72*, 3802–3806.

[13] Momany, F. A., McGuire, R. F., Burgess, A. W. & Scheraga, H. A. (1975) *J. Phys. Chem. 79*, 2361–2381.

[14] Némethy, G., Pottle, M. S. & Scheraga, H. A. (1983) *J. Phys. Chem. 87*, 1883–1887.

[15] Whitney, P. & Tanford, C. (1962) *J. Biol. Chem. 237*, PC1735–PC1737.

[16] Tanford, C. (1970) *Advances in Protein Chemistry 24*, 1–95.

[17] Merlini, G. & Westermark, P. (2004) *J. Intern. Med. 255*, 159–178.

[18] Sunde, M., Serpell, L. C., Bartlam, M., Fraser, P. E., Pepys, M. B. & Blake, C. C. (1997) *J. Mol. Biol. 273*, 729–739.

[19] Koo, E. H., Lansbury, P. T. & Kelly, J. W. (1999) *Proc. Natl. Acad. Sci. USA. 96*, 9989–9990.

[20] Vendruscolo, M., Zurdo, J., MacPhee, C. E. & Dobson, C. M. (2003) *Phil. Trans. R. Soc. A 361*, 1205–1222.

[21] Levinthal, C. (1968) *J. Chim. Phys. 65*, 44–45.

[22] Dinner, A. R., Sali, A., Smith, L. J., Dobson, C. M. & Karplus, M. (2000) *Trends in Biochemical Sciences 25*, 331–339.

[23] Zwanzig, R., Szabo, A. & Bagchi, B. (1992) *Proc. Natl. Acad. Sci. USA. 89*, 20–22.

[24] Karplus, M. (1997) *Folding and Design 2*, S69–S75.

[25] Kim, P. S. & Baldwin, R. L. (1990) *Annual Review of Biochemistry 59*, 631–660.

[26] Matouschek, A., Serrano, L. & Fersht, A. R. (1992) *J. Mol. Biol. 224*, 819–835.

[27] Clarke, D. T., Doig, A. J., Stapley, B. J. & Jones, G. R. (1999) *Proc. Natl. Acad. Sci. USA. 96*, 7232–7237.

[28] Burton, R., Huang, G., Daugherty, M., Calderone, T. & Oas, T. (1997) *Nature Struct. Biol. 4*, 305–310.

[29] Eaton, W. A., Muñoz, V., Thompson, P. A., Henry, E. R. & Hofrichter, J. (1998) *Accounts Chem. Res. 31*, 745–753.

[30] Onuchic, J. N., Luthey-Schulten, Z. & Wolynes, P. G. (1997) *Annu.Rev.Phys.Chem. 48*, 545–600.

[31] Dobson, C. M., Šali, A. & Karplus, M. (1998) *Angew. Chem. Int. Ed. 37*, 869–893.

[32] Bell, S. & Crighton, J. (1984) *J. Chem. Phys. 80*, 2464–2475.

[33] Daggett, V., Li, A., Itzhaki, L., Otzen, D. & Fersht, A. (1996) *J. Mol. Biol. 257*, 430–440,.

[34] Martinez, J. C., Pisabarro, M. T. & Serrano, L. (1998) *Nature Struct. Biol. 5*, 721–729.

[35] Bryngelson, J. D., Onuchic, J. N., Socci, N. D. & Wolynes, P. G. (1995) *Proteins: Structure, Function and Genetics 21*, 167–195.

[36] Dill, K. & Chan, H. (1997) *Nature Struct. Biol. 4*, 10–19.

[37] Goldbeck, R. A., Thomas, Y. G., Chen, E., Esquerra, R. M. & Kliger, D. S. (1999) *Proc. Natl. Acad. Sci. USA. 96*, 2782–2787.

[38] Ladurner, A., Itzhaki, L., Daggett, V. & Fersht, A. (1998) *Proc. Natl. Acad. Sci. USA. 95*, 8473–8478.

[39] Cavalli, A., Haberthür, U., Paci, E. & Caflisch, A. (2003) *Protein Science 12*, 1801–1803.

[40] Verkhivker, G. M., Rejto, P. A., Gelhaar, D. K. & Freer, S. T. (1996) *Proteins: Structure, Function and Genetics 25*, 342–353.

[41] Eaton, W. A., Munoz, V., Thompson, P. A., Chan, C. K. & Hofrichter, J. (1997) *Curr. Opin. Struct. Biol. 7*, 10–14.

[42] Williams, S., Causgrove, T. P., Gilmanshin, R., Fang, K. S., Callender, R. H., Woodruff, W. H. & Dyer, R. B. (1996) *Biochemistry 35*, 691–697.

[43] Shuker, H., Hajduk, P., Meadows, R. & Fesik, S. W. (1996) *Science 274*, 1531–1534.

[44] Dyson, H. J. & Wright, P. E. (1998) *Nature Struct. Biol. 5*, 499–503.

[45] Fersht, A. *Structure and mechanism in protein science: a guide to enzyme catalysis and protein folding.* W. H. Freeman and Co., New York, (1999).

[46] Plaxco, K. W. & Dobson, C. M. (1996) *Curr. Opin. Struct. Biol. 6*, 630–636.

[47] Weiss, S. (1999) *Science 283*, 1676–1683.

[48] Mehta, A. D., Rief, M., Spudich, J. A., Smith, D. A. & Simmons, R. D. (1999) *Science 283*, 1689–1695.

[49] Ishii, Y. & Yanagida, T. (2000) *Single Molecules 1*, 5–13.

[50] Bursulaya, B. D. & Brooks III, C. L. (1999) *J. Am. Chem. Soc. 121*, 9947–9951.

[51] Snow, Y. M., Nguyen, N., Pande, V. & Gruebele, M. (2002) *Nature 42*, 102–106.

[52] Jang, S., Shin, S. & Pak, Y. (2002) *J. Am. Chem. Soc. 124*, 4976–4979.

[53] Ferrara, P. & Caflisch, A. (2000) *Proc. Natl. Acad. Sci. USA. 97*, 10780–10785.

[54] Gsponer, J., Habertür, U. & Caflisch, A. (2003) *Proc. Natl. Acad. Sci. USA. 100*, 5154–5159.

[55] Balbirnie, M., Grothe, R. & Eisenberg, D. (2001) *Proc. Natl. Acad. Sci. USA. 98*, 2375–2380.

[56] Stefani, M. & Dobson, C. M. (2003) *J. Mol. Med 81*, 678–699.

[57] Serpell, L., Sunde, M., Benson, M., Tennent, G., Pepys, M. & Fraser, P. (2000) *J. Mol. Biol. 300*, 1033–1039.

[58] Serpell, L. (2000) *Biochem. Biophys. Acta 1502*, 16–30.

[59] Chamberlain, A., MacPhee, C., Zurdo, J., Morozova-Roche, L., Hill, H., Dobson, C. & J.J., D. (2000) *Biophys. J. 79*, 3282–3293.

[60] Jimenez, J., Nettleton, E., Bouchard, M., Robinson, C., Dobson, C. & Saibil, H. (2002) *Proc. Natl. Acad. Sci. 99*, 9196–9201.

[61] Glenner, G., Ein, D., Eanes, E., Bladen, H., Terry, W. & Page, D. (1971) *Science 174*, 712–714.

[62] Colon, W. Kelly, J. (1992) *Biochemistry 31*, 8654–8660.

[63] Gujiarro, J., Sunde, M., Jones, J., Campbell, I. & Dobson, C. (1998) *Proc. Natl. Acad. Sci. 95*, 4224–4228.

[64] Litvinovich, S., Brew, S., Aota, S., Akiyama, S., Haudenschild, C. & Ingham, K. (1998) *J. Mol. Biol. 280*, 245–258.

[65] Dobson, C. M. (1999) *Trends Biochem. Sci. 24*, 329–332.

[66] Chiti, F., Taddei, N., White, P. M., Bucciantini, M., Magherini, F., Stefani, M. & Dobson, C. M. (1999) *Nature Struct. Biol. 6*, 1005–1009.

[67] Prusiner, S. (1997) *Science 278*, 245–251.

[68] Prusiner, S. B. (1991) *Science 252*, 1515–1522.

[69] Prusiner, S. B. (1996) *TIBS 21*, 482–487.

[70] Morrissey, M. P. & Shakhnovich, E. I. (1999) *Proc. Natl. Acad. Sci. U.S.A. 96*(20), 11293–11298.

[71] Pan, K. M. *et al.* (1993) *Proc. Natl. Acad. Sci. USA. 90*, 10962–10966.

[72] Zhang, H., Kaneko, K., Nguyen, J. T., Livshits, T. L., Baldwin, M. A., Cohen, F. E., James, T. L. & Prusiner, S. B. (1995) *J. Mol. Biol. 250*, 514–526.

[73] Weissmann, C., Fischer, M., Raeber, A., Bueler, H., Sailer, A., Shmerling, D., Rulicke, T., Brandner, S. & Aguzzi, A. (1996) *Cold Spring Harb. Symp. Quant. Biol. 61*, 511–522.

[74] Ma, B. & Nussinov, R. (2002) *Proc. Natl. Acad. Sci. USA. 99*, 14126–14131.

[75] Petkova, A. T., Ishii, Y., Balbach, J. J., Antzutkin, O. N., Leapman, R. D., Delaglio, F. & Tycko, R. (2002) *Proc. Natl. Acad. Sci. USA. 99*(26), 16742–16747.

[76] Tjernberg, L. O., Callaway, D. J. E., Tjernberg, A., Hahne, S., Lilliehöök, C., Terenius, L., Thyberg, J. & Nordstedt, C. (1999) *J. Biol. Chem. 274*(18), 12619–12625.

[77] Malinchik, S. B., Inouye, H., Szumowski, K. E. & Kirschner, D. A. (1998) *Biophys. J. 74*, 537–545.

[78] Citron, M. (2004) *Trends in Pharmacological Sciences 25*, 92–97.

[79] Clarke, G., Collins, R., Leavitt, B., Andrews, D., Hayden, M., Lumsden, C. & McInnes, R. (2000) *Nature 496*, 195–199.

[80] Perutz, M. F. & Windle, A. H. (2001) *Nature 412*, 143–144.

[81] Chiti, F., Taddei, N., baroni, F., Capanni, C., Stefani, M., Ramponi, G. & Dobson, C. (2002) *Nature 9*, 137–143.

[82] Chiti, F., Stefani, M., Taddei, N., Ramponi, G. & Dobson, C. (2003) *Nature 424*, 805–808.

[83] DuBay, K. F., Pawar, A. P., Chiti, F., Zurdo, J., Dobson, C. M. & Vendruscolo, M. (2004) *J. Mol. Biol. 341*, 1317–1326.

[84] Fernandez Escamilla, A. M., Rousseau, F., Schymkowitz, J. & Serrano, L. (2004) *Nat. Biotech. 22*, 1302–1306.

[85] Marcotte, E. M. & Eisenberg, D. (1999) *Biochemistry 38*, 667–676.

[86] Mattews, D. & Cooke, B. (2003) *Rev. sci. tech. Off. int. Epiz. 22*, 283–296.

[87] Hill, A. F., Joiner, S., Linehan, J., Desbruslais, M., Lantos, P. L. & Collinge, J. (2000) *Proc. Natl. Acad. Sci. USA. 97*, 10248–10253.

[88] Dunker, A. K., Brown, C. J., Lawson, J. D., M.Iakoucheva, L. & Obradovic, Z. (2002) *Biochemistry 41*, 6574–6582.

[89] Linding, R., Schymkowitz, J., Rousseau, J., Diella, F. & Serrano, L. (2004) *J. Mol. Biol. 342*, 345–353.

[90] Williams, R. M., Obradovic, Z., Mathura, V., Braun, W., Garner, E. C., Young, J., Takayama, S., Brown, C. J. & Dunker, A. K. (2001) *Pac. Symp. Biocomput. 200*, 89–100.

[91] Si, K., Linquist, S. & Kandel, E. R. (2003) *Cell 115*, 879–891.

[92] Si, K., Giustetto, M., Etkin, A., Hsu, R., Janisiewicz, A. M., Miniaci, M. C., H., J., Zhu, H. & Kandel, E. R. (2003) *Cell 115*, 893–904.

Chapter 2

A Computational Study on the β-Aggregation

Aggregation properties of 1632 peptides are studied by using the genetic algorithm *Lila*, written by G.G. Tartaglia. *Lila* searches the space of sequences for those which have the best match to a certain three-dimensional structure providing results which are largely discussed in Chapter 3, 4, and 5.

2.1 Genetic algorithm optimization

Aggregation propensities of small peptides are investigated with a genetic algorithm optimization in sequence space and molecular dynamics sampling of conformation space. As target structures for the optimizations we used the parallel and the antiparallel β-sheet conformations of three aggregated replicas of the Sup35 yeast prion peptide `GNNQQNY` [1] (Figure 2.1):

- For each peptide sequence, three replicas are submitted to a molecular dynamics simulation starting from the target conformation.
- Peptides sequences are ranked according to their ability to prevent disaggregation using the fitness function. The fitness function for each sequence is estimated to be the number of snapshots whose C_α root mean square deviation (RMSD) from the template is lower than 1 Å.
- The best matches are replicated and subjected to mutations and cross over.

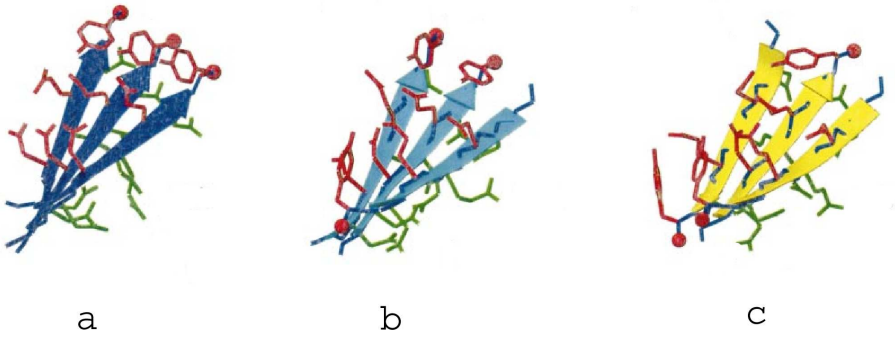

Fig. 2.1: Aggregation of three heptapeptides: *a* parallel, *b* mixed, and *c* antiparallel conformation [1].

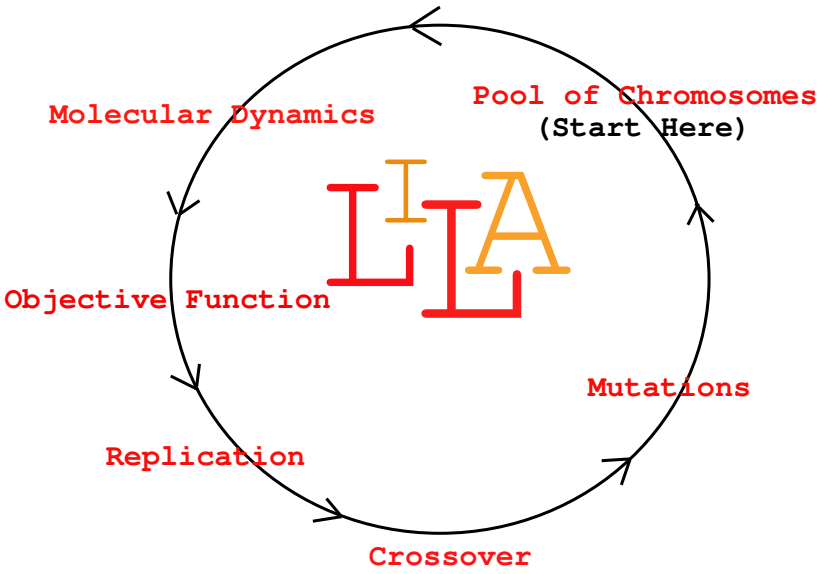

Fig. 2.2: Sketch of the genetic algorithm optimization (*LILA*).

2.2 Molecular dynamics

The *MD* simulations and part of the analysis of the trajectories were performed with the *CHARMM* program [2]. The oligomeric peptide systems were modeled by explicitly considering all heavy atoms and the hydrogen atoms bound to nitrogen or oxygen atoms (PARAM19 potential function [2, 3]). The remaining hydrogen atoms are considered as part of the carbon atoms to which they are covalently bound (extended atom approximation). The effective energy, whose negative gradient corresponds to the force used in the dynamics, is:

$$E(r) = E_{vacuum}(r) + G_{solv}(r)$$

for a molecular system with atomic nuclei located at $r = (r_1, ..., r_N)$. In vacuo, the *PARAM19* energy function is:

$$\begin{aligned} E_{vacuum}(r) = & \frac{1}{2}\sum_{bonds} k_b(b-b_0)^2 + \frac{1}{2}\sum_{bond\ angles} k_\theta(\theta-\theta_0)^2 \\ & + \frac{1}{2}\sum_{dihedral\ angles} k_\phi[1+\cos(n\phi-\delta)] \\ & + \frac{1}{2}\sum_{improper\ dihedrals} k_\omega(\omega-\omega_0)^2 \\ & + \sum_{i>j} \epsilon_{i\ j}^{min}[(\frac{d_{ij}^{min}}{r_{ij}})^{12} - 2(\frac{d_{ij}^{min}}{r_{ij}})^6] \\ & + \sum_{i>j} \frac{q_i q_j}{\epsilon(r_{ij}) r_{ij}} \end{aligned}$$

where b is a bond length, k_θ a bond angle, ϕ a dihedral angle, k_ω an improper dihedral, r_{ij} is the distance between atoms i and j, q_i and q_j are partial charges, and d_{ij}^{min} and ϵ_{ij}^{min} are the optimal van der Waals distance and energy, respectively. An implicit model based on the solvent accessible surface was used to describe the main effects of the aqueous solvent on the solute [4]. In this approximation, the solvation free energy is given by:

$$G_{solv}(r) = \sum_{i=1}^{N} \sigma_i A_i(r)$$

for a molecular system having N heavy atoms with Cartesian coordinates $r = (r_1, ..., r_N)$. $A_i(r)$ is the solvent-accessible surface computed by an

approximate analytical expression and using a 1.4 Å probe radius. The solvation model contains only two σ parameters: one for carbon and sulfur atoms ($\sigma_{C,S} = 0.012$ $kcal/mol$ Å), and one for nitrogen and oxygen atoms ($\sigma_{N,O} = -0.060$ $kcal/mol$ Å). Hydrophobic side chains tend to be buried within the solute whereas hydrophilic side chains and the polar groups of the backbone prefer to be solvent accessible. Furthermore, ionic side chains were neutralized and a linear distance-dependent screening function $\epsilon(r_{ij}) = 2r_{ij}$ was used for the electrostatic interactions. The *PARAM19* default cutoffs for long range interactions were used, i.e., a shift function was employed with a cutoff at 7.5 Åfor both the electrostatic and van der Waals terms. This cutoff length was chosen to be consistent with the parameterization of the force-field and implicit solvation model. The model is not biased toward any particular secondary structure type. In fact, exactly the same force field and implicit solvent model have been used recently in MD simulations of aggregation and folding of structured peptides [5, 6] (α-helices and $\beta-$sheets) ranging in size from 15 to 31 residues [7, 8], and small proteins of about 60 residues [9, 1].

2.3 Lila's set up and performances

In two independent runs of *Lila*, a pool of *48* sequences is subjected to 17 evolutionary cycles: 1632 sequences are generated, for a total amount of $81\,\mu s$ of molecular dynamics simulation (Figure 2.2).

- **First cycle**: Sequences are generated randomly.
- **Molecular dynamics**: Each peptide sequence is submitted to $50\,ns$ molecular dynamic simulation with implicit solvent (see previous section). A non-physiological temperature of *330 K* is used to obtain enough sampling in the time scale of the simulations [1]. In order to optimize the *CPU* time, no periodic boundary conditions are used and peptides are free to separate. To avoid long-distance calculations, we measured the fitness as the number of snapshots whose $C_\alpha - RMSD$ from the template is lower than *1* Å.
- **Best parents**: For each evolutionary cycle *three* best matches are selected. The best parent *nr. 1* has the highest fitness function, followed by the best parent *nr. 2* and the best parent *nr. 3*.
- **Replication**: *50%* of the pool is filled by replicas of the best parent *nr. 1*, *30%* by replicas of the best parent *nr. 2*, and *20%* by replicas of the best parent *nr. 3*.
- **Cross over:** Two by two, all the sequences are subjected to random cross over. One replica of the best parents remains unchanged (*elitism*).

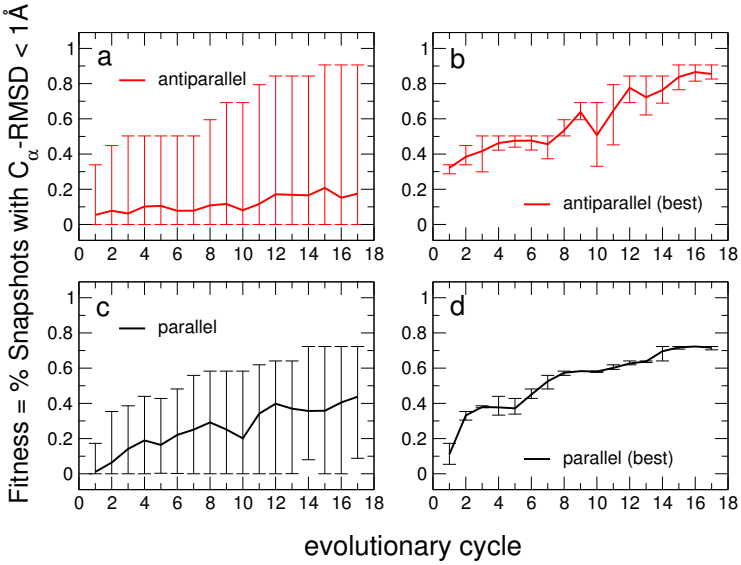

Fig. 2.3: Lila's performances for the parallel and the antiparallel β-sheet aggregation. For a total of 17 cycles, the fitness function is estimated to be the number of snapshots whose C_α-RMSD from the template is lower than 1 Å.

- **Mutations**: All the sequences are subjected to random mutations. For each sequence the mutation involves from zero to three amino acid modifications (assigned randomly).

The performances of *Lila* are reported in Figure 2.3. For each evolutionary cycle, the average of the fitness displays a monotonic trend, which points out the effective optimization of the population. Moreover, the average and the dispersion of the fitness are found proportional (Plots *a* and *c* of Figure 2.3), indicating that there is no premature convergence to non-optimal solutions. Plots *b* and *d* of Figure 2.3 show that the fitness of the best parents reaches the maximum value of 0.7 for the parallel optimization and 0.9 for the antiparallel optimization (Table 2.2).

In the optimizations, amino acid changes follow specific patterns. In sequences selected for the parallel β-sheet aggregation the number of aliphatic and aromatic residues increases almost monotonically, while the number of charged and polar residues decreases. The opposite is observed in sequences selected for the antiparallel β-sheet aggregation (see Figure 2.4). The analysis indicates that parallel aggregates are stabilized by hydrophobic interactions (mainly π-stacking of aromatic residues), while antiparallel

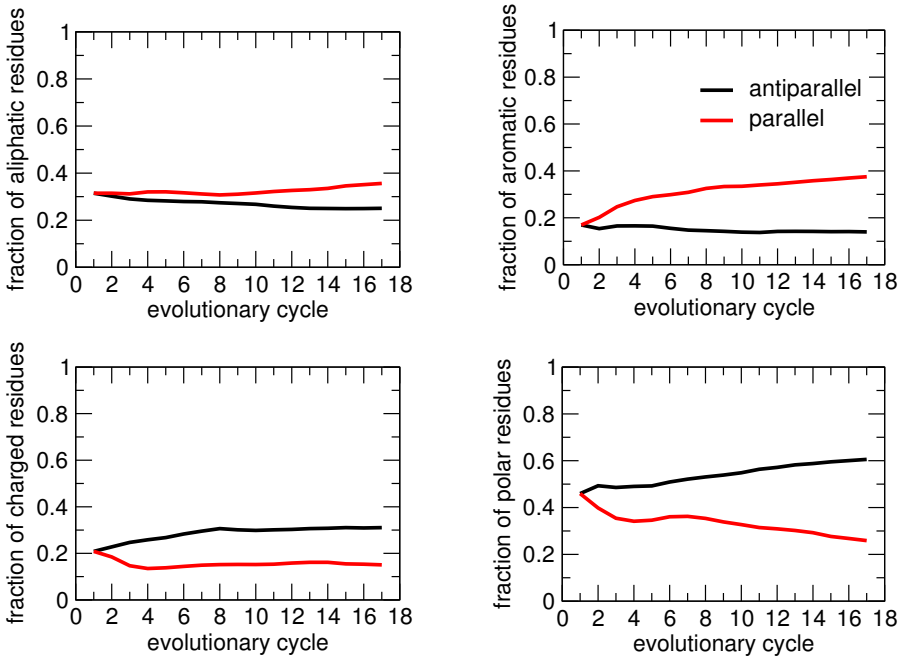

Fig. 2.4: Trends of amino acid properties for the parallel and antiparallel optimizations. In the plots, the number of aliphatic, aromatic, charged, and polar residues is normalized by the length of the peptide and averaged over the population.

aggregates are stabilized by electrostatic interactions (mainly dipole-dipole interactions).

Sequences selected for the parallel β-sheet aggregation show high identity with fragments of the Alzheimer's Aβ_{40} [10]. In particular, we found that *160* sequences have three matches with the fragment `QKLVFFA` and *20* sequences have four matches with `HQKLVFF` which is also known to be amyloidogenic [11]. We also found the sequences `HFWLVFF` and `FFVLYQH` which display five and six inverted matches with with the fragment `HQKLVFF`. By considering that the genetic algorithm sampled 816 sequences during the optimization of the parallel aggregation and a random search approximately needs 10^7 sequences to scan before finding six matches, we conclude that the genetic algorithm approach performs 10^4 better than random. Compared to known amyloidogenic fragments [12], sequences selected for the antiparallel aggregation show no significant matches.

Sequence^a	**Fitness**	**Sequence**^{p}	**Fitness**
YNTIVDF	0.33	LQYQMLY	0.17
YATIDRY	0.44	YEWLKRY	0.33
KVTCDRY	0.50	YVWYKFY	0.37
KVTCDRY	0.50	LWYQMYY	0.42
KVTCDRY	0.50	LWYQKFY	0.48
KVTCDRY	0.50	AWYQKFY	0.55
KVTCDRY	0.50	YAWYKFY	0.58
KVTSNVY	0.59	YAWLKFY	0.58
KDTQDRY	0.69	YAWLKFY	0.58
KDTQDRY	0.69	YAWLKFY	0.58
YDCQDFY	0.79	YEWLKFY	0.61
TDTQDFE	0.84	YFWLKFY	0.64
TDTQDFE	0.84	YFWLKFY	0.64
TDTQDFE	0.84	MFWLYFY	0.72
TDTCDWQ	0.90	MFWLYFY	0.72
TDTCDWQ	0.90	MFWLYFY	0.72
TDTCDWQ	0.90	MFWLYFY	0.72

Tab. 2.2: Sequence and fitness of best parents *nr. 1*. The labels a and p indicate peptide sequences whose optimized β-aggregated conformation is antiparallel or parallel, respectively.

2.4 Fibril formation

Peptide sequences are optimized by *Lila* to prevent disaggregation. In order to test the aggregation propensity of best parents, we performed molecular dynamics simulations of the peptides `YDCQDFY`, `TDTQDFE`, and `TDTCDWQ` (selected for the antiparallel β-sheet aggregation) as well as of the peptides `YEWLKFY`, `YFWLKFY`, and `MFWLKFY` (selected for the parallel β-sheet aggregation). For each peptide, three replicas are randomly placed in a cubic box of $75X75X75\text{Å}^3$ and submitted to a $330\ K$ molecular dynamics simulation of $1\,\mu s$ using periodic boundary conditions. Figure 2.5 displays the histogram of Π, i.e., the number of snapshots whose C_α-RMSD from the parallel or antiparallel target structure is lower than 1 Å. Except for `YDCQDFY`, all the best parents show a main peak in corrispondence of $\Pi \sim 1$, which indicates the predominance of ordered aggregates. As expected, the peptides `YEWLKFY`, `YFWLKFY`, and `MFWLKFY` form parallel aggregates, while the peptides `TDTQDFE` and `TDTCDWQ` form antiparallel aggregates. The peptides `TDTQDFE` and `TDTCDWQ` display a peak in corrispondence of $\Pi \sim 0.5$, which indicates the presence of a mixed parallel-antiparallel conformation (Figure 2.1), due to the π-stacking of aromatic residues [13]. Among the tested best parents, the peptide `YFWLKFY` shows the highest tendency to the ordered aggregation.

Six replicas of the peptide `YFWLKFY` are submitted to a molecular dynamics simulation in the conditions explained above. Since no reference structure is available for the fibril, we used two different variables to monitor the aggregation progress: The orientation parameter P_2 [14] and the number of parallel and antiparallel C_α-contacts of the peptides (Figure 2.6). The P_2 parameter (defined in the range $[0, 1]$) indicates that all the peptides take the same orientation, while the number of C_α-contacts (defined in the range $[0, 35]$) displays the predominance of parallel contacts. The P_2 parameter reaches a maximum at the value of 0.92, which points out the presence of a fibril twist (Figure 2.7). Both the variables indicate that the six replicas form a parallel β-sheet aggregate in $2.5\,\mu s$, which is the stricking consequence of the genetic algorithm optimization.

2.5 β-Aggregation Matrices

The β-aggregation propensity of sequences is investigated by using a procedure which was previously developed to estimate the energy of proteins from the primary structure alone [15, 16]. Our approach allows the measure of side-chains contributions to the β-aggregation and can be applied to compute the aggregation propensity of other sequences.

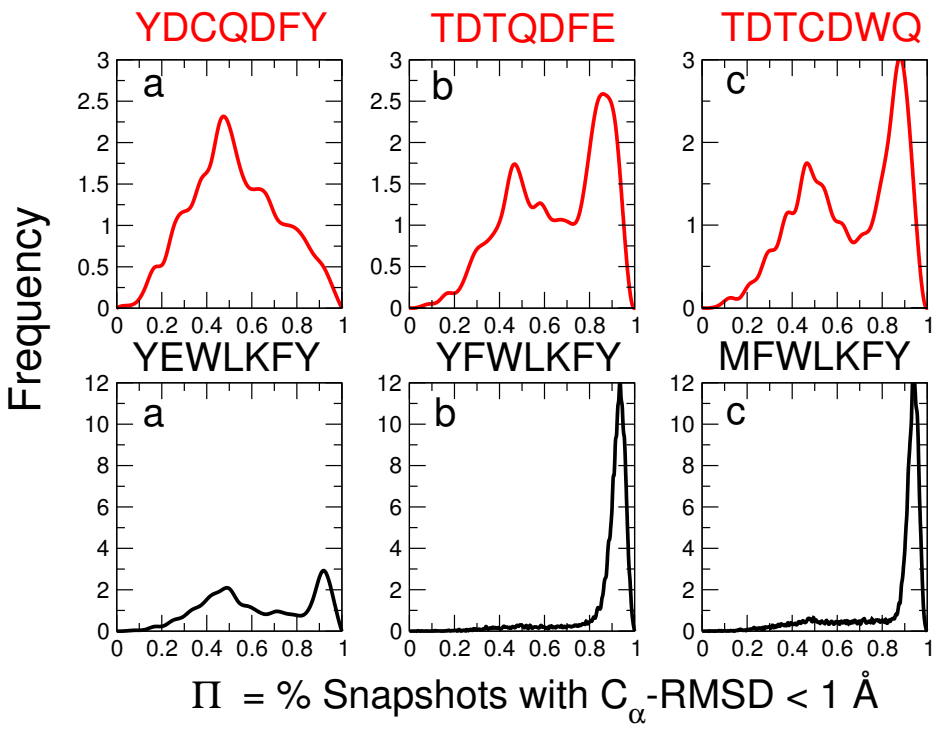

Fig. 2.5: Aggregation simulation of best parents. The variable Π indicates the number of snapshots whose C_α-RMSD from the parallel (black) or antiparallel (red) target structure is lower than 1 Å is used to build the histogram.

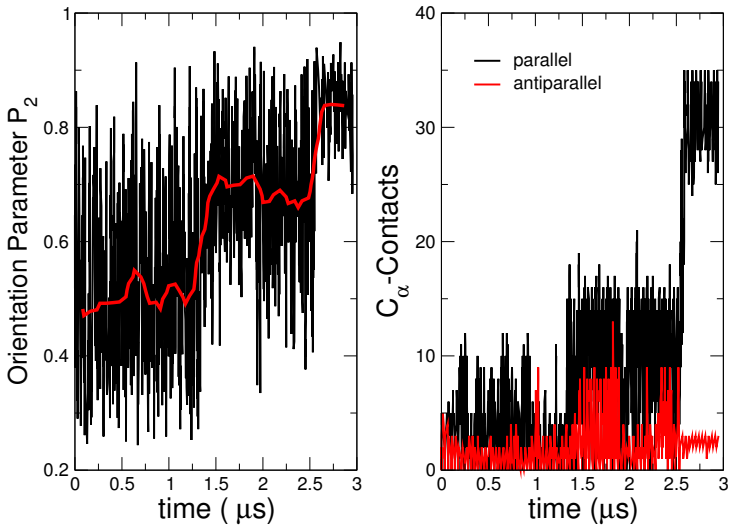

Fig. 2.6: Aggregation of six replica-peptides `YFWLKFY`. Two variables are used to monitor the aggregation progress: The orientation parameter P_2 (defined in the range $[0, 1]$) and the number of C_α-contacts between peptides (defined in the range $[0, 35]$).

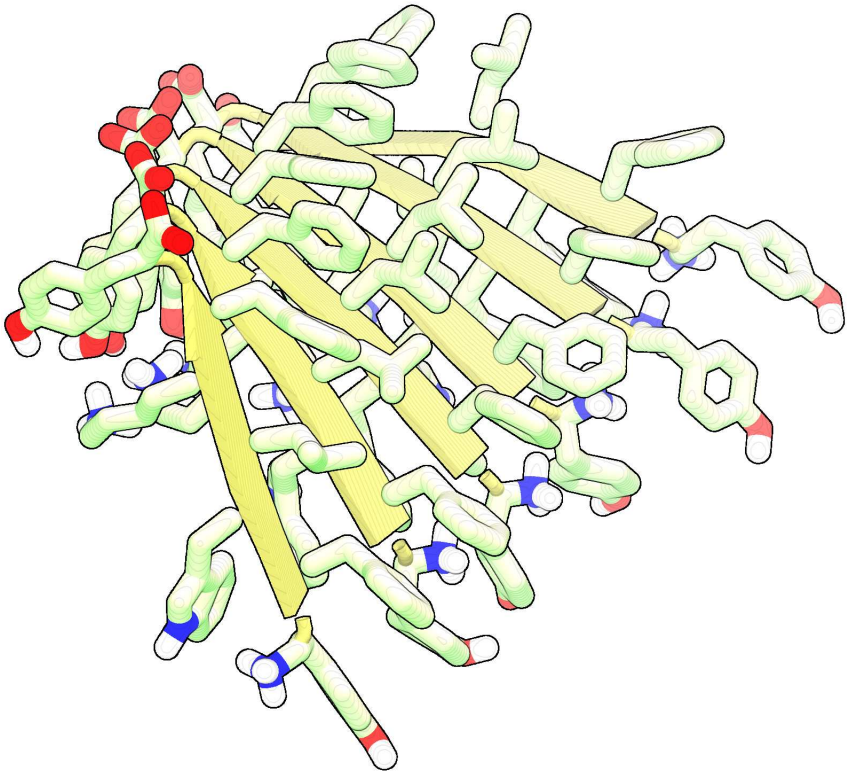

Fig. 2.7: Six replica peptides `YFWLKFY` forming a twisted fibril.

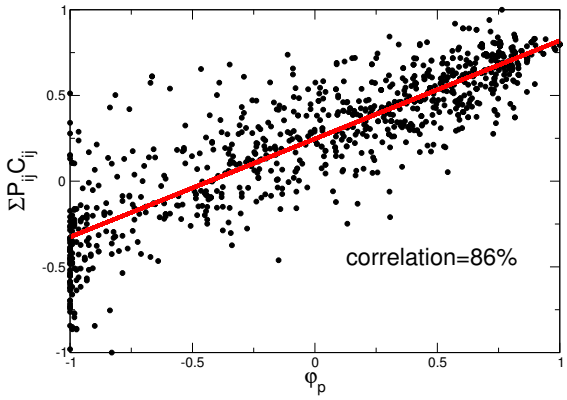

Fig. 2.8: Parallel aggregation: The normalized fitness φ_p is used to train the β-aggregation propensity matrix P

For computational reasons, the fitness f is transformed into the variable φ, normalized in the interval $[-1, 1]$:

$$\varphi = 2\left(\frac{1}{2} - \frac{f - f_{max}}{f_{min} - f_{max}}\right)$$

where f_{min} and f_{max} are the lowest and the highest values of the fitness, respectively. A value of φ close to 1 means $f \sim f_{max}$, while a value of φ close to -1 means $f \sim f_{min}$. The variables φ_p and φ_a, corresponding to the fitness for the parallel and the antiparallel aggregation, are used to compute the matrices P and A:

$$\varphi_p \overset{C_p}{\rightarrow} P$$

$$\varphi_a \overset{C_a}{\rightarrow} A$$

where C_p and C_a are matrices which describe the side-chains contacts in the aggregated parallel and antiparallel β-sheet conformation, respectively. The two $20X20$ symmetric matrices P and A describe the parallel and the antiparallel β-aggregation propensity of the amino acid sequence:

$$\varphi_p \simeq \sum P_{ij} C_p^{ij}$$

$$\varphi_a \simeq \sum A_{ij} C_a^{ij}$$

Figures 2.8 and 2.9 show that φ_p and φ_a train the matrices P and A with a correlation of 75% and 86%, respectively. The high correlations

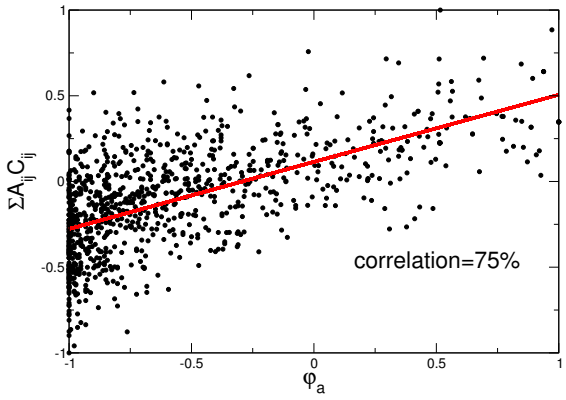

Fig. 2.9: Antiparallel aggregation: The normalized fitness φ_a is used to train the β-aggregation propensity matrix A.

found suggest that fitness values are not affected by statistical errors due to insufficient molecular dynamics sampling and that sequences generated by genetic algorithm optimization contain a signal which is largely superior to noise. The matrix A shows lower correlation than the matrix B because of the high number of low-fitness sequences ($\varphi_a \sim -1$). In fact, the presence of charged residues causes electrostatic interactions which are unfavorable for several amino acid mutations.

The eigenvalues of A and P, sorted from the largest negative value λ^1 to the largest positive value λ^{20}, were used to determine the eigenvectors $\underline{v}$ and write the diagonal form of φ_p and φ_a:

$$\varphi_p \sim \lambda_p^1 \sum_{i,j} v_i^1(p) C_p^{ij} v_j^1(p) + \lambda_p^2 \sum_{i,j} v_i^2(p) C_p^{ij} v_j^2(p) + ... + \lambda_p^{20} \sum_{i,j} v_i^{20}(p) C_p^{ij} v_j^{20}(p)$$

$$\varphi_a \sim \lambda_a^1 \sum_{i,j} v_i^1(a) C_a^{ij} v_j^1(a) + \lambda_a^2 \sum_{i,j} v_i^2(a) C_a^{ij} v_j^2(a) + ... + \lambda_a^{20} \sum_{i,j} v_i^{20}(a) C_a^{ij} v_j^{20}(a)$$

Since $\sum_{i,j} v_i^1 C^{ij} v_j^1 > 0$, the sign of the eigenvalue λ determines when the corresponding eigenvector stabilizes ($\lambda > 0$) or destabilizes ($\lambda < 0$) the β-aggregation. The eigenvector $\underline{v}^{20}$ provides the description of the most stabilizing contributions. For the antiparallel configuration, the eigenvector $\underline{v}^{20}(a)$ shows that charged residues stabilize aggregation (Figure 2.10) and displays a correlation of 70% with Eisenberg's hydrophobicity scale [17]. For the parallel configuration, the eigenvector $\underline{v}^{20}(p)$ indicates that aromatic

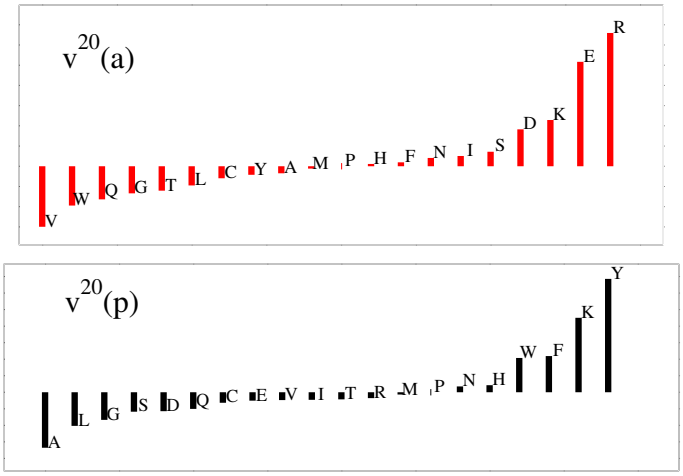

Fig. 2.10: Eigenvector analysis. The two eigenvectors v_a^{20} and v_p^{20} respectively correspond to the largest eigenvalues of the matrices A and P and indicate that charged residues stabilize the antiparallel configuration and aromatic residues stabilize the parallel configuration.

residues stabilize the parallel aggregation (Figure 2.10) and displays a correlation of 70% with Mayo's β-propensity scale [18]. We plan in the future to use the matrices A and P to design new sequences.

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	-2.57	-1.97	-1.63	1.71	-5.33	0.82	0.69	-0.45	-14.76	1.71	-9.20	0.21	0.00	-1.75	-0.72	-2.08	-1.55	-2.61	-3.66	-5.18
C	-1.97	0.11	0.32	-0.02	-1.76	-0.10	0.20	-0.79	-3.04	-3.23	-1.91	1.48	0.00	-3.35	0.52	-0.83	-0.59	-0.10	-1.02	-2.79
D	-1.63	0.32	-0.87	0.91	-2.47	0.77	0.19	0.16	-1.97	-1.09	-1.45	0.63	0.00	1.61	-0.81	0.68	1.41	0.43	-0.02	-3.86
E	1.71	-0.02	0.91	-1.34	2.03	-1.54	0.27	0.07	-0.44	1.52	0.07	-0.27	0.00	-0.81	-0.22	3.02	0.53	0.62	3.62	-2.38
F	-5.33	-1.76	-2.47	2.03	4.63	-1.29	0.60	-1.29	4.12	2.44	4.69	1.22	0.00	3.23	2.84	-2.20	0.98	-1.84	-9.89	6.46
G	0.82	-0.10	0.27	-1.54	-1.29	-0.18	0.45	-0.23	-2.25	-7.41	-0.12	-0.20	0.00	-2.12	-0.06	-0.47	-1.54	-0.33	-1.27	-7.96
H	0.69	0.20	0.37	0.94	0.60	0.45	-0.73	0.56	-1.01	0.29	-0.64	0.41	0.00	-0.72	0.15	0.91	0.75	0.08	1.22	2.93
I	-0.45	-0.79	0.16	0.97	-1.29	-0.23	0.56	-1.07	-1.77	0.98	-0.27	-0.06	0.00	0.67	0.20	0.21	0.03	-0.05	3.18	-1.44
K	-14.76	-3.04	-1.97	-0.44	4.12	-2.25	-1.01	-1.77	-11.34	-10.09	-12.51	-2.65	0.00	-6.51	-5.33	-5.09	-1.14	-1.78	17.20	8.03
L	1.71	-3.23	-1.09	1.52	2.44	-7.41	0.29	0.98	-10.09	-2.81	0.46	-1.95	0.00	-8.91	0.30	-2.55	2.21	-2.98	8.52	-10.01
M	-9.20	-1.91	-1.45	0.07	4.69	-0.12	-0.64	-0.27	-12.51	0.46	-2.06	-1.31	0.00	-3.10	-0.40	-3.03	-3.53	-3.02	7.42	-2.34
N	0.21	1.48	0.63	-0.27	1.22	-0.20	0.41	-0.06	-2.65	-1.95	-1.31	-0.88	0.00	0.85	0.02	-1.23	0.57	0.09	0.28	2.56
P	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Q	-1.75	-3.35	1.61	-0.81	3.23	-2.12	-0.72	0.67	-6.51	-8.91	-3.10	0.85	0.00	-0.79	-0.25	-0.91	-0.67	0.09	2.19	-6.55
R	-0.72	0.52	-0.81	-0.22	2.84	-0.06	0.15	0.20	-5.33	0.30	-0.40	0.02	0.00	-0.25	-0.71	0.74	-0.20	-0.63	3.56	-0.53
S	-2.08	-0.83	0.68	3.02	-2.20	-0.47	0.91	0.21	-5.09	-2.55	-3.03	-1.23	0.00	-0.91	0.02	-0.73	0.54	0.84	0.87	-3.53
T	-1.55	-0.59	1.41	0.53	0.98	-1.54	0.75	0.03	-1.14	2.21	-3.53	0.57	0.00	-0.67	-0.20	0.54	-2.40	0.96	4.70	-3.48
V	-2.61	-0.10	0.43	0.62	-1.84	-0.33	0.08	-0.05	-1.78	-2.98	-3.02	0.73	0.00	0.09	-0.63	0.84	0.96	-0.25	0.69	-2.72
W	-3.66	-1.02	-0.02	3.62	-9.89	-1.27	1.22	3.18	17.20	8.52	7.42	0.28	0.00	2.19	3.56	0.87	4.70	0.69	-9.82	6.36
Y	-5.18	-2.79	-3.86	-2.38	6.46	-7.96	2.93	-1.44	8.03	-10.01	-2.34	2.56	0.00	-6.55	-0.53	-3.53	-3.48	-2.72	6.36	11.69

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	0.06	-3.32	-1.21	-1.34	-2.49	-1.42	0.68	4.23	-0.66	0.27	1.22	-0.92	0.00	-1.64	-1.26	-0.92	-2.17	-1.12	-1.09	-1.81
C	-3.32	4.82	-3.39	1.65	-3.39	-2.52	-0.05	0.70	-2.94	-2.29	-2.52	-2.84	0.00	-0.17	-1.37	-5.54	-8.53	1.99	-2.63	1.14
D	-1.21	-3.39	-1.32	-0.30	2.84	-5.56	1.11	2.69	1.90	4.04	3.63	0.20	0.00	13.47	7.47	3.83	9.81	-2.40	6.18	10.50
E	-1.34	1.65	-0.30	-1.54	-0.36	-1.23	1.04	-1.44	2.30	-0.31	-0.80	-1.10	0.00	-0.79	5.14	-2.01	2.74	-3.87	-3.92	-2.79
F	-2.49	-3.39	2.84	-0.36	-1.70	-1.63	-0.64	-1.64	0.95	-1.11	1.89	0.21	0.00	0.93	-2.57	-1.32	-1.47	-2.16	-0.69	-3.46
G	-1.42	-2.52	-5.56	-1.23	-1.63	-0.98	-0.49	-0.39	-0.05	-0.68	0.75	-0.27	0.00	-1.95	-2.43	-0.54	-3.08	-2.77	-0.87	-1.25
H	0.68	-0.05	1.11	1.04	-0.64	-0.49	0.43	-0.24	-1.55	-0.23	-1.52	-0.18	0.00	1.30	-0.96	-1.02	-0.31	-1.32	-0.76	0.05
I	4.23	0.70	2.69	-1.44	-1.64	-0.39	-0.24	0.18	0.45	0.74	0.18	1.30	0.00	-0.60	1.49	-1.41	2.18	0.99	-1.20	-3.62
K	-0.66	-2.94	1.90	2.30	0.95	-0.05	-1.55	0.45	-3.37	2.60	-0.25	0.38	0.00	-4.44	-0.88	0.88	-6.72	-0.23	-2.18	3.65
L	0.27	-2.29	4.04	-0.31	-1.11	-0.68	-0.23	0.74	2.60	-1.54	-0.72	-0.49	0.00	0.14	-3.07	-1.00	1.54	0.41	-0.38	1.39
M	1.22	-2.52	3.63	-0.80	1.89	0.75	-1.52	0.18	-0.25	-0.72	0.47	1.38	0.00	1.35	-0.73	-1.29	-2.34	0.63	-0.73	-2.47
N	-0.92	-2.84	0.20	-1.10	0.21	-0.27	-0.18	1.30	0.38	-0.49	1.38	0.42	0.00	0.66	-0.56	-1.40	-0.80	-1.62	-2.73	1.42
P	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Q	-1.64	-0.17	13.47	-0.79	0.93	-1.95	1.30	-0.60	-4.44	0.14	1.35	0.66	0.00	0.41	-3.48	-3.68	4.35	0.55	-2.66	2.00
R	-1.26	-1.37	7.47	5.14	-2.57	-2.43	-0.96	1.49	-0.88	-3.07	-0.73	-0.56	0.00	-3.48	-0.48	-0.08	-4.70	0.58	-3.93	-3.70
S	-0.92	-5.54	3.83	-2.01	-1.32	-0.54	-1.02	-1.41	0.88	-1.00	-1.29	-1.40	0.00	-3.68	-0.08	-1.59	0.61	2.15	-3.15	-2.28
T	-2.17	-8.53	9.81	2.74	-1.47	-3.08	-0.31	2.18	-6.72	1.54	-2.34	-0.80	0.00	4.35	-4.70	0.61	-0.90	2.33	-2.87	-2.29
V	-1.12	1.99	-2.40	-3.87	-2.16	-2.77	-1.32	0.99	-0.23	0.41	0.63	-1.62	0.00	0.55	0.58	2.15	2.33	4.65	-3.58	-1.74
W	-1.09	-2.63	6.18	-3.92	-0.69	-0.87	-0.76	-1.20	-2.18	-0.38	-0.73	-2.73	0.00	-2.66	-3.93	-3.15	-2.87	-3.58	-1.75	-1.35
Y	-1.81	1.14	10.50	-2.79	-3.46	-1.25	0.05	-3.62	3.65	1.39	-2.47	1.42	0.00	2.00	-3.70	-2.28	-2.29	-1.74	-1.35	-7.08

Tab. 2.3: Parallel (*top*) and antiparallel (*bottom*) β-sheet propensity.

Bibliography

[1] Gsponer, J. & Caflisch, A. (2001) *J. Mol. Biol. 309*, 285–298.

[2] Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. & Karplus, M. (1983) *J. Comput. Chem. 4*, 187–217.

[3] Neria, E., Fischer, S. & Karplus, M. (1996) *J. Chem. Phys. 105*, 1902–1921.

[4] Ferrara, P., Apostolakis, J. & Caflisch, A. (2002) *Proteins: Structure, Function and Genetics 46*, 24–33.

[5] Paci, E., Gsponer, J., Salvatella, X. & Vendruscolo, M. (2004) *J. Mol. Biol. 340*, 555.

[6] Gsponer, J., Habertür, U. & Caflisch, A. (2003) *Proc. Natl. Acad. Sci. USA. 100*, 5154–5159.

[7] Hiltpold, A., Ferrara, P., Gsponer, J. & Caflisch, A. (2000) *J. Phys. Chem. B 104*, 10080–10086.

[8] Ferrara, P. & Caflisch, A. (2001) *J. Mol. Biol. 306*, 837–850.

[9] Gsponer, J. & Caflisch, A. (2002) *Proc. Natl. Acad. Sci. USA. 99*, 6719–6724.

[10] Tjernberg, L. O., Callaway, D. J. E., Tjernberg, A., Hahne, S., Lilliehöök, C., Terenius, L., Thyberg, J. & Nordstedt, C. (1999) *J. Biol. Chem. 274*(18), 12619–12625.

[11] Williams, A. D., Portelius, E., Kheterpal, I., Guo, J., Cook, K. D., Xu, Y. & Wetzel, R. (2004) *J. Mol. Biol. 335*, 833–842.

[12] Tartaglia, G. G., Cavalli, A., Pellarin, R. & Caflisch, A. (2005) *Protein Science in press* .

[13] Gazit, E. (2002) *FASEB J. 16*, 77–83.

[14] Harrison, P. M., Chan, H. S., Prusiner, S. B. & Cohen, F. E. (1999) *J. Mol. Biol. 286*, 593–606.

[15] Dosztanyi, Z., Csizmok, V., Tompa, P. & Simon, I. (2005) *J. Mol. Biol. 347*, 827–839.

[16] Thomas, P. D. & Dill, K. A. (1996) *Proc. Natl Acad. Sci. 93*, 11628–11633.

[17] Eisenberg, D., Schwarz, E., Komaromy, M. & Wall, R. (1984) *J. Mol. Biol. 179*, 125–142.

[18] Street, A. & Mayo, S. (1999) *Proc. Natl. Acad. Sci. USA 96*, 9074–9076.

Chapter 3

The Role of Aromaticity, Exposed Surface, and Dipole Moment in Determining Protein Aggregation Rates

Protein Science (2004) 13, 1939-1941

FOR THE RECORD

The role of aromaticity, exposed surface, and dipole moment in determining protein aggregation rates

GIAN GAETANO TARTAGLIA, ANDREA CAVALLI, RICCARDO PELLARIN, AND AMEDEO CAFLISCH

Department of Biochemistry, University of Zurich, CH-8057, Zurich, Switzerland

(RECEIVED February 2, 2004; FINAL REVISION March 18, 2004; ACCEPTED March 25, 2004)

Abstract

The mechanisms by which peptides and proteins form ordered aggregates are not well understood. Here we focus on the physicochemical properties of amino acids that favor ordered aggregation and suggest a parameter-free model that is able to predict the change of aggregation rates over a large set of natural sequences. Furthermore, the results of the parameter-free model correlate well with the aggregation propensities of a set of peptides designed by computer simulations.

Keywords: amyloid; prion; aggregation rate; Alzheimer; protein deposit; mutation

Supplemental material: see www.proteinscience.org

Amyloid fibrils are involved in a number of diseases, including Alzheimer's disease, Parkinson's disease, Huntington's disease, prion disease, and type II diabetes (Kelly 1998; Rochet and Lansbury Jr. 2000). Therefore, it is of fundamental medical interest to understand the mechanisms of fibrillogenesis with the ultimate goal of designing inhibitors. The amyloid fibril formation is not a property limited to a selected few proteins: Under certain conditions it has been shown that any polypeptide chain can form fibrils (Dobson 1999). Because aggregation conditions vary sensibly with the composition and sequence of the polypeptide, single amino acid substitution has been used to investigate the fibril formation (Chiti et al. 2002). In this study we propose a formula to predict the change of aggregation and disaggregation rate upon mutation. The agreement between the experimental data and our formula leads us to the conclusion that the formation of fibrils can be explained with a simple model based on physicochemical properties of amino acids. We found that the polar and the nonpolar water-accessible surface areas, the dipole moment, and the π-stacking interaction of aromatic residues (Gazit 2002) are essential beside the charge and the β-propensity of the sequence (Chiti et al. 2003). To have the most possible general model, we do not use any parameter that needs to be experimentally estimated. Furthermore, our equation does not present any redundancy, whereas in previous work by others charge and hydrophobicity were considered independent and used as two different variables in the best-fitting (Chiti et al. 2003).

We propose the following function to predict the effect of a mutation on aggregation rate:

$$\nu_{mut}/\nu_{wt} = \phi_h\phi_\beta\phi_a\phi_c \quad (1)$$

where ν_{wt} and ν_{mut} are the aggregation rates of the wild type and mutant, respectively. The factor ϕ_h captures most of the nonpolar and polar interactions. An amino acid is called *p* if its side chain carries a charge or a dipole; otherwise it is called *a*.

For mutations that involve same type of amino acids $a \rightarrow a$ or $p \rightarrow p$

$$\phi_h^I = \begin{cases} ASA^a_{mut}/ASA^a_{wt} & a \rightarrow a \\ ASA^p_{wt}/ASA^p_{mut} & p \rightarrow p \end{cases}$$

Reprint requests to: Amedeo Caflisch, Department of Biochemistry, University of Zurich, CH-8057, Zurich, Switzerland; e-mail: caflisch@bioc.unizh.ch; fax: +41-44-635-68-62.

Article published online ahead of print. Article and publication date are at http://www.proteinscience.org/cgi/doi/10.1110/ps.04663504.

where ASA^a and ASA^p are the nonpolar and polar water-accessible surface areas of the amino acid side chains (Makhatadze and Privalov 1990; Karplus 1997). Interestingly, experimental evidence has been published recently on the importance of nonpolar solvent-accessible surface area for the amyloid-like properties of apomyoglobin (Chow et al. 2003).

For mutations that involve different types of amino acids ($a \rightarrow p$ or $p \rightarrow a$)

$$\phi_h^{II} = \begin{cases} 1/D_{mut} & a \rightarrow p \\ D_{wt} & p \rightarrow a \end{cases}$$

where D is the magnitude of the dipole of the amino acid side chains. The function ϕ_h^I implies that the hydrophobicity and aggregation rate increase as the mutation results in a larger nonpolar surface or smaller polar surface. In ϕ_h^{II}, it has been assumed that the nonpolar surface of p amino acids compensates the nonpolar surfaces of a amino acids so that the dipole of p amino acids exclusively characterizes the mutation (see Supplementary Table 1).

The factor ϕ_β is related to the ratio of β-sheet propensities (Street and Mayo 1999; see Supplementary Table 1):

$$\phi_\beta = \frac{\beta_{mut}}{\beta_{wt}}$$

Functions ϕ_a and ϕ_c approximate the effect of the aromatic residues A and total charge C, respectively:

$$\phi_a \phi_c = e^{\Delta A} e^{-\Delta |C|/2}$$

The factor ½ before C has been introduced to have the same range [−1, 1] for the arguments of the two exponential functions.

In Figure 1 our model is used to predict the changes in aggregation rates occurring in human muscle acylphosphatase (AcP), islet amyloid polypeptide, prion peptides, α-synuclein, amyloid β-peptide, tau, leucine-rich repeat, and some model peptides. As in Chiti et al. (2003), we divided the data set in two parts to compare with their equation. The correlation obtained with equation 1 is significant (85% and 86% and $P < 10^{-4}$), and slightly better than the one obtained by Chiti et al. using three parameters derived from best fitting (76% and 85% and $P < 10^{-4}$). The good agreement with experiments shows that our simple equation, which does not contain any parameter, is very general and can be used to describe the aggregation of several and heterogeneous protein systems.

The validity of the formula is proved also by rearranging the whole data set per a and p mutations: Slopes and correlations are very close (see Supplementary Fig. 1; $p \rightarrow p$: slope = 1.01, correlation = 80%, number of points = 28; $a \rightarrow a$: slope = 0.92, correlation = 82%, number of points = 15; $a \rightarrow p$ and $p \rightarrow a$: slope = 1.01, correlation = 89%, number of points = 12).

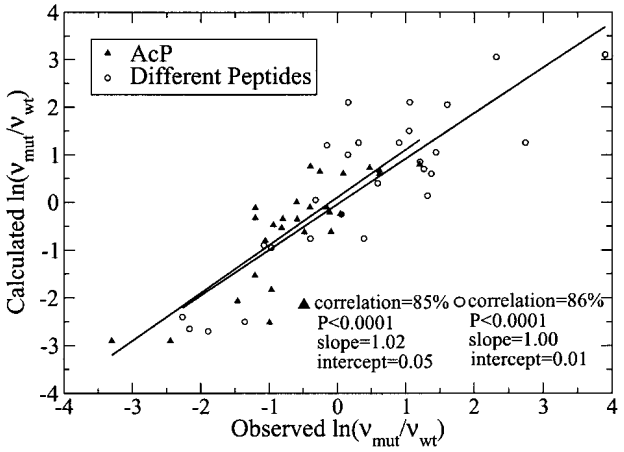

Figure 1. Calculated vs. observed (Chiti et al. 2003) changes in aggregation rate upon mutation: AcP (28 triangles) and heterogeneous groups of peptide and protein systems, including islet amyloid polypeptide, prion peptides, α-synuclein, amyloid β-peptide, τ, leucine-rich repeat and some model peptides (27 circles).

Aggregation and disaggregation are intrinsically different, but the role played by the hydrophobicity, β-propensity, π-stacking, and charge is the same. Considering that disaggregation and aggregation are opposite processes, the direct proportionality relation between ν_{mut}/ν_{wt} and $\phi_h\phi_\beta\phi_a\phi_c$ that describes the aggregation turns into a relation of inverse proportionality for the disaggregation. Therefore, the reciprocal of equation 1 can be used to describe the disaggregation:

$$\nu_{wt}/\nu_{mut} = \phi_h\phi_\beta\phi_a\phi_c \quad (2)$$

To verify the validity of this assumption, we applied equation 2 to heptapeptide sequences suggested by a genetic algorithm approach (G. Tartaglia and A. Caflisch, in prep.). The genetic algorithm searches the space of sequences for those that have the best match to a certain three-dimensional target conformation (an in-register parallel aggregate of three heptapeptides [Gsponer et al. 2003]). For each peptide sequence, three replicas are submitted to a 330 K molecular dynamics simulation, starting from the β-parallel aggregated conformation (CHARMM parameter 19 [Brooks et al. 1983] and solvent accessible surface-based solvation model [Ferrara et al. 2002]). A temperature of 330 K is used to obtain enough sampling in the time scale of the simulations (Gsponer et al. 2003). Peptide sequences are ranked according to their ability to prevent disaggregation. The disaggregation rate is estimated for each sequence as the reciprocal of the number of snapshots whose C_α root mean square deviation (RMSD) from the template is lower than 1 Å. Best

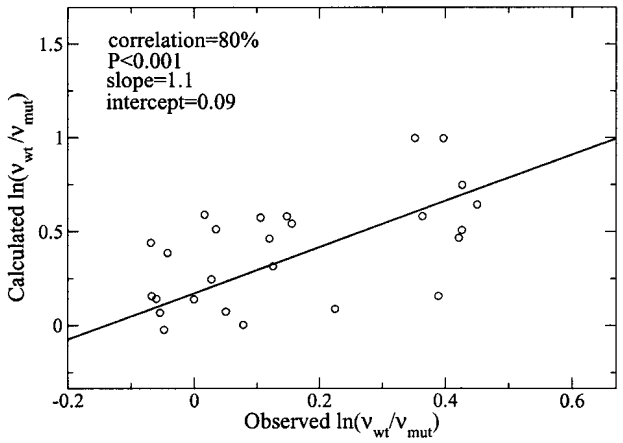

Figure 2. Calculated vs. observed changes in disaggregation rate upon mutation: Best parents of genetic algorithm approach (27 circles). (See Supplementary Table 2.)

matches, called best parents, are replicated and subjected to mutations and crossover: 10^3 sequences have been studied for a total amount of 50 μsec of simulation. The genetic algorithm predicted several sequences similar to segments of amyloidogenic protein as well as the sequence HFWLVFF, which presents five matches with the amyloid β-peptide fragment HQKLVFF (Tjernberg et al. 1999; Williams et al. 2004). By considering that the genetic algorithm sampled 10^3 sequences and a random search approximately needs 10^6 sequences to scan before finding five matches, we conclude that the genetic algorithm approach performs 10^3 better than random.

Disaggregation rates are analyzed with equation 2 only for best parents (4% of data) for which false positives are supposed to be less than the false negatives in the remaining set. Furthermore, to have statistical significance, each disaggregation rate has been averaged over a set of five molecular dynamics trajectories. Figure 2 shows that equation 2 holds and the correlation is very high (80% and $P < 10^{-3}$). In conclusion, the present results indicate that a simple model based on physicochemical properties without parametrization is able to predict aggregation and disaggregation rates.

Acknowledgments

We thank Dr. E. Paci for interesting discussions and Prof. F. Chiti for providing rates of AcP. This work was supported by the Swiss National Science Foundation (grant no. 31-64968.01 to A.C.) and the National Center of Competence in Research (NCCR) in Structural Biology.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

References

Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S., and Karplus, M. 1983. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **4:** 187–217.

Chiti, F., Calamai, M., Taddei, N., Stefani, M., Ramponi, G., and Dobson, C. 2002. Studies of the aggregation of mutant proteins in vitro provide insights into the genetics of amyloid diseases. *Proc. Natl. Acad. Sci.* **99:** 16419–16426.

Chiti, F., Stefani, M., Taddei, N., Ramponi, G., and Dobson, C. 2003. Rationalization of the effects of mutations on peptide and protein aggregation rates. *Nature* **424:** 805–808.

Chow, C., Chow, C., Raghunathan, V., Huppert, T.J., Kimball, E.B., and Cavagnero, S. 2003. Chain length dependence of apomyoglobin folding: Structural evolution from misfolded sheets to native helices. *Biochemistry* **42:** 7090–7099.

Dobson, C.M. 1999. Protein misfolding, evolution and disease. *Trends Biochem. Sci.* **24:** 329–332.

Ferrara, P., Apostolakis, J., and Caflisch, A. 2002. Evaluation of a fast implicit solvent model for molecular dynamics simulations. *Proteins* **46:** 24–33.

Gazit, E. 2002. A possible role for π-stacking in the self-assembly of amyloid fibrils. *FASEB J.* **16:** 77–83.

Gsponer, J., Haberthür, U., and Caflisch, A. 2003. The role of side-chain interactions in the early steps of aggregation: Molecular dynamics simulations of an amyloid-forming peptide from the yeast prion Sup35. *Proc. Natl. Acad. Sci.* **100:** 5154–5159.

Karplus, P. 1997. Hydrophobicity regained. *Protein Sci.* **6:** 1302–1307.

Kelly, J. 1998. The alternative conformations of amyloidogenic proteins and their multi-step assembly pathways. *Curr. Opin. Struct. Biol.* **8:** 101–106.

Makhatadze, G. and Privalov, P. 1990. Heat capacity of proteins 1 partial molar heat capacity of individual amino acid residues in aqueous solution: Hydration effect. *J. Mol. Biol.* **213:** 375–384.

Rochet, J.C. and Lansbury Jr., P.T. 2000. Amyloid fibrillogenesis: Themes and variations. *Curr. Opin. Struct. Biol.* **10:** 60–68.

Street, A. and Mayo, S. 1999. Intrinsic β-sheet propensities result from van der Waals interactions between side chains and the local backbone. *Proc. Natl. Acad. Sci.* **96:** 9074–9076.

Tjernberg, L., Callaway, D., Tjernberg, A., Hahne, S., Lilliehook, C., Terenius, L., Thyberg, J., and Nordstedt, C. 1999. A molecular model of Alzheimer amyloid β-peptide fibril formation. *J. Biol. Chem.* **274:** 12619–12625.

Williams, A., Portelius, E., Kheterpal, I., Guo, J., Cook, K., Xu, Y., and Wetzel, R. 2004. Mapping A_β amyloid fibril secondary structure using scanning proline mutagenesis. *J. Mol. Biol.* **335:** 833–842.

Chapter 4

Prediction of Aggregation Rate and Aggregation-prone segments in Polypeptide Sequences

Protein Science (2005) 14, 2723-2734

Prediction of aggregation rate and aggregation-prone segments in polypeptide sequences

GIAN GAETANO TARTAGLIA, ANDREA CAVALLI, RICCARDO PELLARIN, AND AMEDEO CAFLISCH

Department of Biochemistry, University of Zürich, CH-8057 Zürich, Switzerland

(RECEIVED March 23, 2005; FINAL REVISION June 23, 2005; ACCEPTED July 4, 2005)

Abstract

The reliable identification of β-aggregating stretches in protein sequences is essential for the development of therapeutic agents for Alzheimer's and Parkinson's diseases, as well as other pathological conditions associated with protein deposition. Here, a model based on physicochemical properties and computational design of β-aggregating peptide sequences is shown to be able to predict the aggregation rate over a large set of natural polypeptide sequences. Furthermore, the model identifies aggregation-prone fragments within proteins and predicts the parallel or anti-parallel β-sheet organization in fibrils. The model recognizes different β-aggregating segments in mammalian and nonmammalian prion proteins, providing insights into the species barrier for the transmission of the prion disease.

Keywords: Alzheimer's disease; amyloid; protein aggregation rate; prion protein; species barrier; genetic algorithm; molecular dynamics

Amyloid fibrils are associated with a number of pathologies including Alzheimer's, Parkinson's, Huntington's, prion disease, and type II diabetes (Horwich and Weissman 1997; Kelly 1998; Dobson 1999; Rochet and Lansbury 2000). Therefore it is of fundamental medical interest to understand the mechanisms of fibrillogenesis, with the ultimate goal of designing inhibitors. One important and still unanswered question regarding amyloid fibril formation is the specificity with which the amino acid sequence determines β-aggregation propensity and the atomic details of the fibril structure. Because of the difficulties in obtaining detailed structural information by X-ray crystallography or solution phase NMR spectroscopy, computational approaches are needed to guide experiments, e.g., to determine short segments of amyloid-like proteins that share the same biophysical properties of the full-length proteins (Balbirnie et al. 2001) and identify those elements which are essential for the formation of protein fibrils (Tenidis et al. 2000; von Bergen et al. 2000). As aggregation conditions vary sensibly with the composition and especially the sequence of the polypeptide, single amino acid substitutions have been used to investigate the fibril formation (Chiti et al. 1999), and complementary theoretical studies proposed relative rate equations to predict the change of aggregation rate upon mutation (Chiti et al. 2003; Tartaglia et al. 2004). Although the application of relative rate equations shows high correlation with experimental data, these models require the a priori knowledge of wild-type aggregation rates.

We report here an absolute rate equation derived from both first principles and analysis of aggregating sequences designed by a computational approach. The latter is based on a genetic algorithm optimization in sequence space and molecular dynamics sampling of conformation space. The equation does not need any information except the amino acid sequence and two environmental factors (i.e., temperature and concentration). Our model gives both the aggregation rate and the "amyloid spectrum" of a protein, identifying those segments involved in β-aggregation. In

Reprint requests to: Amedeo Caflisch, Department of Biochemistry, University of Zürich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland; e-mail: caflisch@bioc.unizh.ch; fax: +41-44-635-68-62.

Article and publication are at http://www.proteinscience.org/cgi/doi/10.1110/ps.051471205.

addition, the model distinguishes between the parallel and anti-parallel β-sheet organization within the fibrils and shows that mammalian and nonmammalian prion proteins have different amyloid spectra.

Results and Discussion

Absolute rate prediction

Predicted and experimentally measured rates are shown in logarithmic scale in Figure 1. The correlation is 95% and extends over 90 data points and about 15 natural logarithmic units. This is a remarkable result considering that the rate is calculated solely from the primary structure with the addition of two external factors, i.e., temperature and concentration. Interestingly, the correlation is good for different proteins and also within mutants of the same protein. For single-point mutants of long sequences (Acylphosphatase and Titin), the error is rather large because of the poor signal-to-noise ratio due to the average over the entire sequence. The model was subjected to statistical tests to assess the chance correlation. In Figure 2A, the experimentally measured rates were randomly permutated to generate about 10^7 "scrambled" data sets. The calculated rates were fitted to each scrambled set, giving an extremely small likelihood for high correlations. In Figure 2B, $\sim 10^7$ data sets were randomly generated within the range of experimental rates. The predictive ability and correlation of the model are much higher than the corresponding values obtained upon randomization of the experimental rates. These statistical tests show that chance correlation is not present.

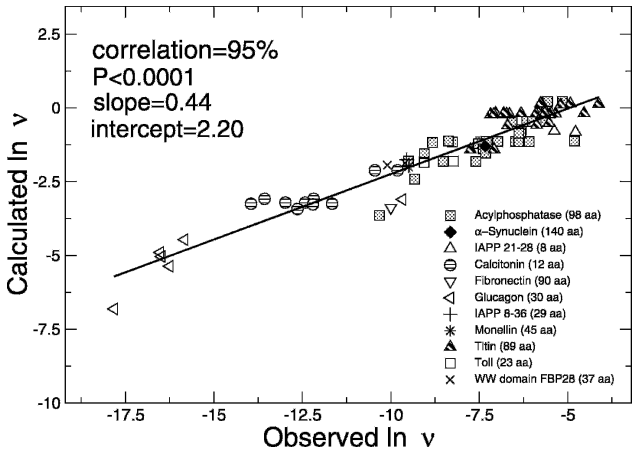

Figure 1. Calculated (Equation 4; see Materials and Methods) vs. observed aggregation rates for heterogeneous groups of peptide and protein systems (Litvinovich et al. 1998; Konno et al. 1999; Chiti et al. 2003; Ferguson et al. 2003; DuBay et al. 2004). A t-student test on the correlation shows the high significance in the prediction (in the present study $P < 0.0001$, while $P \simeq 1$ indicates no significance).

Prediction of β-aggregating segments

There is in vivo evidence that amyloid fibrils originate from misfunctions of the degradation machinery and cleavage of fragments that have high propensity for β-aggregation (Stefani and Dobson 2003). Moreover, even proteins not implicated in amyloid diseases were recently found to form amyloid fibrils in vitro under denaturing conditions, indicating that fibrillogenesis is a common feature of proteins (Chiti et al. 1999; Dobson 1999; Stefani and Dobson 2003). Our approach to estimate aggregation rates can be also used to identify segments with high aggregation propensity. The method is tested on the following proteins: α-synuclein, apolipoprotein, amyloid precursor protein (APP), gelsolin, islet amyloid precursor protein (IAPP), lactadherin, prion, serum amyloid A, transthyretin, ABri, ADan, fibrinogen, β_2-microglobulin, insulin, Sup35, and tau. The former nine proteins represent all hits of a combined search for "amyloid" and "human" at http://www.expasy.org (Gasteiger et al. 2003) in September 2004; the latter seven proteins result from a literature search (references are reported in Table 1). As indicated in Figure 3, the data set contains

- regions known to promote aggregation
- segments found to aggregate in vivo (often after degradation)
- stretches extracted from the precursors and shown to aggregate in vitro

Each sequence in the data set is scanned by shifting a window of fixed size one residue at a time starting from the N terminus. The extracted stretches are ranked using the aggregation propensity π (see Materials and Methods). The procedure is repeated for different window sizes (3–25 amino acids), each time storing the positions of the three stretches having the highest π. These positions are then used to build the histogram of Figure 3. Peaks of the histogram represent positions of stretches with the highest β-aggregation propensity ("windows' consensus"). All the sequences except fibrinogen and prion show main peaks in segments known to promote aggregation. For prion, amyloidogenic areas are—up to now—not known and few experiments have been performed and on limited portions of the protein (Vanik et al. 2004). Following the protein-only hypothesis (Prusiner 1988; Soto and Castilla 2004), we suggest that the peak found at position 150 may be determinant for prion transmissions (in the subsection Prions, the same peak is numbered with 175 because of the alignment with other prion sequences). For transthyretin, only one of the two experimentally known β-aggregating fragments has been found with our analysis. We speculate that the corresponding area promotes the aggregation of the entire protein, which is consistent with NMR data (Jaroniec et al. 2002).

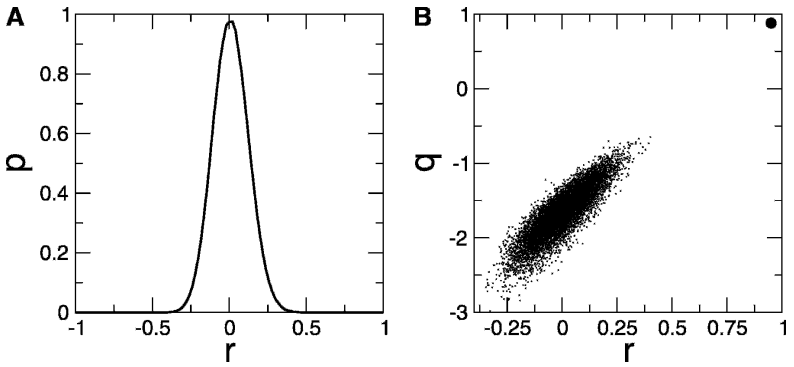

Figure 2. Statistical tests to assess chance correlation. (*A*) Permutations of experimental rates: Probability distribution p of the correlation coefficient r between rates calculated with Equation 4 (see Materials and Methods) and scrambled experimental rates. The likelihood of obtaining high correlations ($r > 50\%$) with scrambled experimental rates is extremely small ($p < 10^{-9}$). (*B*) Randomization of experimental rates (within the same range of values): Cross-validated leave-one-out correlation coefficient $q = 1 - \text{PRESS}/\sigma^2$ (PRESS = predicted residual sum of squares, i.e., sum of squared differences between predicted and observed values [Zoete et al. 2003]) vs. the correlation coefficient r. The predictive ability and correlation of the model (thick circle on the *top right*) are significantly separated from the corresponding values obtained upon randomization of the experimental rates (thin points). In both tests, 10^7 data sets were generated.

To further test the sensitivity of our model, we focused on the segments that are experimentally known to aggregate. For this purpose, we used a window size of five consecutive residues, as in a previous work (Fernandez Escamilla et al. 2004) (Table 1). Interestingly, several five-residue stretches are found in segments that were shown to aggregate, e.g., FGAIL contained in IAPP NFGAILSS, FILDL in gelsolin's SFNNGDCFILD, SVQFV in lacthaderin's NFGSVQFV, and YQQYN in Sup35's PQGGYQQYN (Azriel and Gazit 2001). For APP, three stretches are found in correspondence of the segment LVFFA, which is known to be involved in the aggregation of $A\beta_{40}$ (Williams et al. 2004) (see subsection Amyloid Protein Precursor). Importantly, all the stretches are ranked among those having the

Table 1. *Analysis of experimentally known β-aggregating segments*

Protein	1st Stretch[a]	Rank[b]	2nd Stretch[a]	Rank[b]	3rd Stretch[a]	Rank[b]	Segment	Total length	Ref.
ABri	22{CSRTV}^{a}	5	21{ICRST}^{a}	8	20{LICSR}^{a}	10	1–34	34	El-Agnaf et al. 2001
ADan	22{CFLNF}^{p}	1	23{FNLFL}^{p}	2	24{NLFLN}^{p}	3	1–34	34	El-Agnaf et al. 2004
α-Synuclein	41{EQVTN}^{a}	6	67[SIAAA]^p	12	71[ATGFV]^p	15	41–74	120	Ueda et al. 1993
Apolipoprotein A–I	18[YVDVL]^p	1	28{DYVSQ}^{a}	2	85[EMSKD]^a	3	1–83	242	Nichols et al. 1988
APP	671{LVFFA}^{p}	1	670{KLVFF}^{p}	2	672{VFFAE}^{p}	3	655–696	750	Weidemann et al. 1989
β-Microglobulin	61{SFYLL}^{p}	1	63{TLLYY}^{p}	2	66{YYTEF}^{p}	3	59–79	99	Jones et al. 2003
Fibrinogen	494[FPGFF]^p	7	493[TFPGF]^p	13	482[AAFFD]^p	32	482–504	623	Asl et al. 1997
Gelsolin	187{DCFIL}^{p}	15	188{CFILD}^{p}	23	189{FILDL}^{p}	31	173–243	755	Kangas et al. 1996
IAPP	22{FGAIL}^{p}	1	21{NFGAI}^{p}	2	28[SNTYG]^a	4	1–38	38	Westermark et al. 1987
Insulin	78{ENYCN}^{a}	1	23{RGFFY}^{p}	3	15{ALYLV}^{p}	4	1–38	86	Jimenez et al. 2002
Lactadherin	260{YGNDQ}^{a}	3	259{SYGND}^{a}	4	289[SVQFV]^p	5	245–294	364	Haggqvist et al. 1999
Prion	116{IIHFG}^{p}	1	115{PIIHF}^{p}	2	99[VVGGL]^p	3	1–121	208	Vanik et al. 2004
Serum amyloid A	3{FFSFL}^{p}	2	4{FSFLG}^{p}	3	5{SFLGE}^{p}	4	2–12	104	Westermark et al. 1992
Sup35	77[YQQYN]^a	1	44[YQNYQ]^a	2	67[YQQQY]^a	3	1–112	683	King et al. 1997
Tau	621{SVQIV}^{p}	23	632{SKVTS}^{a}	24	627{KPVDL}^{p}	25	617–636	757	Margittai and Langen 2004
Transthyretin	107[IAALL]^p	1	114{YSYST}^{a}	2	106[TIAAL]^p	4	105–115	127	Jaroniec et al. 2002

[a] The three five-residue stretches with the highest π, within the segments listed in the third to last column, are reported with the predicted parallel (*p*) or anti-parallel (*a*) arrangement. The braces { } indicate stretches that are close to the peak found in the experimental regions using the windows' consensus (Figure 3), while the brackets [ ] mark sequences that are distant from the peak. The integer before the brackets refers to the position of the stretch in the processed protein (initial signal- and pro-peptides are omitted in the notation as in other works; see, for instance Kangas et al. 1996; Jones et al. 2003).

[b] The rank of the stretches refers to the entire precursor protein and can in principle vary from 1 (i.e., the stretch has the highest π among all the stretches in the precursor protein) to the total length of the precursor protein (i.e., the stretch has the lowest π among all the stretches in the precursor protein).

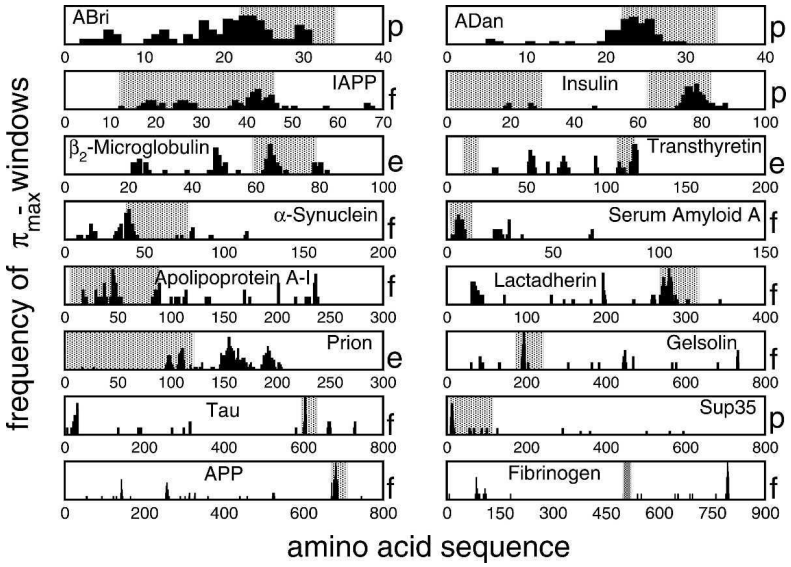

Figure 3. Windows' consensus. Different window sizes (3–25 amino acids) are used to scan proteins. Positions of stretches with highest aggregation propensity π are used to build the histogram. Except for fibrinogen and prion, the highest peak is located in segments that are known to form amyloid fibrils and/or contribute to protein aggregation (gray regions). The letter "p" labels regions that are known to promote fibrillogenesis ("p" standing for "promoting"). The letter "f" indicates segments that are found to aggregate in vivo ("f" standing for "fragment") after degradation. The letter "e" refers to stretches that are shown to aggregate in vitro ("e" standing for "extracted"). We stress that Equation 1 (see Materials and Methods) was used to identify β-aggregating stretches and not to predict amino acid deletions or insertions involved in amyloidosis. Positions refer to proteins without signal- and pro-peptides. References for all the experiments are reported in Table 1.

highest π in the respective precursor proteins (see Table 1), which suggests that a small window size is sufficient for the identification of amyloidogenic regions. In Table 1, we also list β-aggregating segments that have not yet been investigated with experiments in vitro (e.g., YVDVL in apolipoprotein A–I and ENYCN in insulin) and indicate the predicted parallel or anti-parallel arrangement of the individual segments in the fibril.

Amyloid protein precursor

Using a window size of five residues, the amyloid spectrum of the 750-residue APP (Fig. 4) shows a predominant peak at position 671 for the stretch LVFFA. Furthermore, the predicted β-aggregating stretches AIIGL and IGLMV are consistent with solid-state NMR (Antzutkin et al. 2002; Bond et al. 2003) and scanning proline mutagenesis (Williams et al. 2004). The stretches with the highest rate for each window size in the range 3–25 are shown in Table 2 for $A\beta_{42}$. Most of the high-aggregation stretches contain the segment LVFFA and are parallel. As in experiments (Gordon et al. 2004), the segment KLVFFAE has a preferential anti-parallel arrangement, while $A\beta_{42}$ is parallel (Antzutkin et al. 2000; Torok et al. 2002). As shown in clinical reports and oligomerization experiments performed with photo-induced cross-linking of unmodified proteins (Bitan et al. 2003), we found that $A\beta_{42}$ has a higher aggregation propensity than $A\beta_{40}$ ($\ln \pi_{A\beta_{42}} = -7$, $\ln \pi_{A\beta_{40}} = -9$). Interestingly, the experimental evidence indicates that the Ile_{41}–Ala_{42} extension of the 1–40 segment affects the rate of amyloid formation rather than the fibril stability (Jarrett et al. 1993).

Prions

To further investigate the usefulness of our model, the amyloidogenic propensities of the prion protein from different organisms were evaluated using a moving window of five residues along the entire sequence. To compare the amyloid spectra, prion sequences have been aligned using ClustalW (Thompson et al. 1994). It is remarkable that prion sequences in mammals show a peak at position 175 corresponding to the segment SNQNN in human prion (Fig. 5; Table 3; all the notations used to number stretches refer to the major prion proteins, i.e., signal- and/or propeptides are omitted). Such a peak is absent in the chicken and the turtle. Interestingly, the peak is located in a glutamine/asparagine-rich region, which shows high propensity to self-propagate in amyloid fibrils (Michelitsch and Weissman 2000). Other peaks correspond to β-strand 2 (segment NQVYY, conserved in mammals and nonmammals and mutated in NRVYY in chicken) and helix 1 of

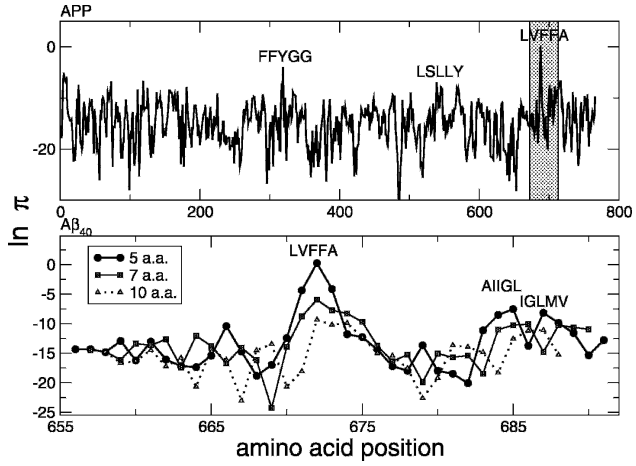

Figure 4. Amyloid protein precursor. The aggregation propensity π is averaged over a window of five amino acids. The entire sequence is scanned by shifting the window by one residue at a time starting from the N terminus ("amyloid spectrum"). The analysis shows a major peak corresponding to the segment LVFFA at position 671. The *bottom* plot focuses on the most amyloidogenic region, which is highlighted in gray in the *top* plot. Windows of different sizes (5, 7, and 10 amino acids), shifted to the central amino acid, give similar results, indicating the robustness of the model. Furthermore, with longer window sizes, peaks in the C terminus of $A\beta_{40}$ become comparable to the one at position 671 (see also Table 2). In both plots, the effective height of the peak is compressed by the logarithm scale.

human prion (segment YEDRY in mammals, WNENS in turtle, and WSENS in chicken), which are known to form ordered aggregates in vitro (Nguyen et al. 1995; Kozin et al. 2001). Furthermore, the amyloid profiles are similar within mammals (e.g., 97% correlation between man and cow) and different between mammals and nonmammals (e.g., 55% correlation between man and turtle).

To compare with experiments in vitro (Vanik et al. 2004), we analyzed the unstructured region of the prion protein (residues 1–122) in human, mouse, and hamster prion peptides. We found that human and mouse prions share similar amyloid spectra (i.e., 98% correlation), while the hamster prion diverges from them at position 143 (position 116 in the nonaligned human sequence). More specifically, the stretch 143–148 of hamster prion (position 116–121 in the nonaligned human sequence) is found to be less amyloidogenic than the corresponding segment in mouse and human ($\ln \pi_{hamster} = -16$, $\ln \pi_{mouse} = -12$, and $\ln \pi_{human} = -12$), which is consistent with the prion 1–122 species barrier observed in vitro (Vanik et al. 2004).

Huntingtin

The gene for Huntington's disease consists of 67 hexons and contains an open reading frame for a polypeptide of > 3140 residues. Using a window size of five residues, our model identifies the N-terminal poly(Gln) repeat and the stretch IFFFL in the middle of the sequence as the two most prone to induce ordered aggregates. With window sizes larger than 20, the N-terminal poly(Gln) repeat dominates and the peak in the middle of the sequence disappears.

Our model is not sensitive enough to discriminate repeats of fewer than 38 glutamine residues from those with > 41 glutamine residues; the former are harmless, whereas the latter are responsible for toxic aggregates (Perutz et al. 1994; Perutz 1999). Alternatively, the dramatic difference in toxicity observed at a repeat length of ~40 might require the context of a much longer polypeptide sequence.

Conclusions

The model presented here was motivated by the challenging tasks of predicting aggregation propensity and identifying β-aggregating stretches in polypeptide sequences. An essential element in the derivation of the equation was the analysis of a large pool of β-aggregating peptide sequences designed by a computational approach based on molecular dynamics and genetic algorithm optimization in sequence space (G.G. Tartaglia and A. Caflisch, in prep.). The very

Table 2. *Stretches of $A\beta_{42}$ with the highest rate at each window size in the range 3–25*

Sequence		ln π	p/a
VFF	{IGL}	5.3 {−2.6}	p
LVFF	{GAII}	2.5 {−6.7}	p
LVFFA	{AIIGL}	0.2 {−7.5}	p
LVFFAE	{GAIIGL}	−3.9 {−8.0}	p
KLVFFAE	{AIIGLMV}	−5.9 {−10.0}	a
LVFFAEDV	{IGLMVGGM}	−7.3 {−10.1}	p
LVFFAEDVG	{GLMVGGVVI}	−7.6 {−10.0}	p
QKLVFFAEDV	{IGLMVGGVVI}	−9.3 {−9.7}	a
{QKLVFFAEDVG}	IGLMVGGVVIA	−10.1 {−11.0}	p
{HQKLVFFAEDVG}	AIIGLMVGGVVI	−10.5 {−11.1}	p
{FFAEDV . . . }	GAIIGLMVGGVVI	−10.5 {−10.7}	p
FFAEDVGSNKGAII		−10.1	p
VFFAEDVGSNKGAII		−9.3	p
VFFAEDVGSNKGAIIG		−9.7	p
LVFFAEDVGSNKGAIIG		−8.8	p
LVFFAEDVGSNKGAIIGL		−8.2	p
KLVFFAEDVGSNKGAIIGL		−9.3	p
KLVFFAEDVGSNKGAIIGLM		−9.4	p
QKLVFFAEDVGSNKGAIIGLM		−10.5	p
QKLVFFAEDVGSNKGAIIGLMV		−10.1	p
LVFFAEDVGSNKGAIIGLMVGGV		−10.7	p
LVFFAEDVGSNKGAIIGLMVGGVV		−10.4	p
LVFFAEDVGSNKGAIIGLMVGGVVI		−7.1	p

In braces are reported stretches that ranked after the highest rate ones and do not overlap with them. The last column contains the preferred β-sheet arrangement, i.e., parallel (p) or anti-parallel (a).

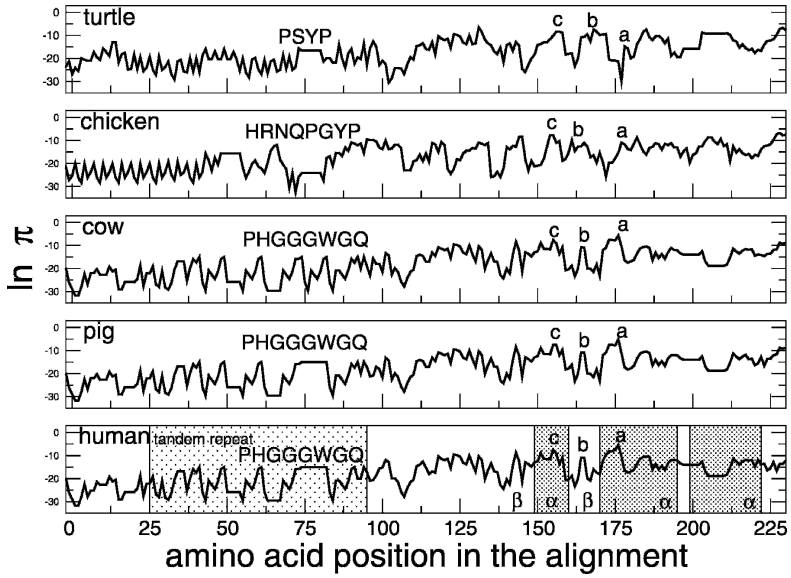

Figure 5. Prion proteins from turtle to human. The plot shows an evolutionary differentiation of the aggregation peaks. Prions of cow and mouse, as well as prions of sheep and pig, show similar amyloid spectra (data not shown). The highest peak at position 175 for mammals (segment *a*, i.e., SNQNN) is not present in nonmammals. Peak *b* (segment NQVYY, conserved in mammals and nonmammals, and mutated to NRVYY in chicken) appears in correspondence of β-strand 2 in human prion. Nonmammals show a peak *c* (segment WNENS in turtle and WSENS in chicken) in correspondence of the first helix of human prion that is weaker in mammals (YEDRY). Sequences have been aligned using ClustalW (Thompson et al. 1994) at http://www.expasy.org/cgi-bin/hub (Gasteiger et al. 2003). Horizontal traits in the plots represent gaps and are meant to help the eye. For all the species, no significant peak is found in the N-terminal tandem repeats. The secondary structural elements of the human prion are labeled with Greek letters and the stretches corresponding to the three α-helices are emphasized by shadowed rectangles.

good correlation between calculated and experimental rates for a large and heterogeneous set of polypeptide chains has allowed us to use the model to successfully identify β-aggregating segments and predict the parallel or anti-parallel arrangement. Fibrils formed by short segments of a protein might have a different molecular structure than the fibril of the full-length protein. Yet our results, as well as previous experimental (Chiti et al. 1999, 2003; Balbirnie et al. 2001) and computational (Fernandez Escamilla et al. 2004) works by others, indicate that the amyloid-forming part of a protein could be only a short segment of the entire chain. That a function based on simple physicochemical principles is able to predict aggregation rates and identify β-aggregating fragments in proteins might be a consequence of the essential role of side-chain interactions in β-sheet aggregates (Gazit 2002; Gsponer et al. 2003; Linding et al. 2004).

Although some of the physicochemical properties in our model are similar to those used in previous works by others, it is important to distinguish approaches based on parameter optimization for a multiterm equation (Chiti et al. 2003; DuBay et al. 2004) from first-principle models like the one of this work and that of Tartaglia et al. (2004). On a very similar test set of peptides and proteins, the multi-parameter approach gives results comparable to those obtained with our model, but it is likely to have a lower predictive ability. As an example, positional effects are taken into account in our model, whereas they are neglected in the multiparameter approach (DuBay et al. 2004), which is mainly based on amino acid composition and alternation of hydrophobic–hydrophilic residues (Broome and Hecht 2000). Recent scanning proline mutagenesis, combined with critical concentration analysis and NMR hydrogen–deu-

Table 3. *Peak at position 175. Prion compatibilies of animals with respect to human*

Animal	$\Delta\pi/\pi$
Turtle	9.52
Chicken	8.72
Sheep	1.66
Pig	1.13
Cow	0.76
Mouse	0.76
Hamster	0.15

The distance with respect to the human prion sequence is measured as $\Delta\pi/\pi = (\pi_{animal} - \pi_{human})/\pi_{human}$ using a window size of five amino acids for the rate calculation and summing over the segment 165–185 to better sample the variability around the peak.

terium exchange, indicate a strong positional effect on both the aggregation kinetics and structural properties of the $A\beta_{40}$ fibril (Williams et al. 2004). Most importantly, the multiparameter approach cannot be used to identify β-aggregating segments as explicitly mentioned by the investigators (DuBay et al. 2004).

Recently, an approach based on secondary structure propensity and estimation of desolvation penalty (TANGO) has been shown to accurately predict the sequence-dependent and mutational effects on the aggregation of a large data set of peptides and proteins (Fernandez Escamilla et al. 2004). TANGO is based on the assumption that the probability of finding > 2 ordered segments in the same polypeptide is negligible. The investigators report that TANGO allows quantitative comparison within the same polypeptide chain or mutants. On the other hand, only qualitative comparison between different polypeptide chains is possible with TANGO (Fernandez Escamilla et al. 2004), whereas our model allows for the prediction of absolute rates (Fig. 1).

In conclusion, we have identified the physicochemical properties of amino acids that are essential for ordered aggregation and proposed a model that takes into account sequence effects for aromatic and charged residues, as well as composition. Compared with the models previously published by others, our equation is the only one that takes explicitly into account π-stacking. Very recent high-resolution structural data (electron and X-ray diffraction) have provided strong evidence for the importance of aromatic side chains for amyloid formation (Makin et al. 2005).

Our model derived from first principles and analysis of *in silico* designed sequences is able to predict aggregation rates and identify β-aggregating segments with high accuracy, suggesting possible biological implications as in the prion protein case. For nonmammalian prions, the absence of the peak at position 175 observed in mammals decreases the overall aggregation propensity, indicating a species-specific behavior consistent with experiments (Marcotte and Eisenberg 1999; Matthews and Cooke 2003) and supporting the hypothesis of a species barrier in the transmission of the prion disease (Hill et al. 2000).

In the accompanying article we present a bioinformatics application of our model that reveals an anti-correlation between organism complexity and proteomic β-aggregation propensity (Tartaglia et al. 2005).

Materials and methods

Absolute rate equation

An equation based on physicochemical properties of natural amino acids is introduced to estimate the aggregation rate of proteins and identify β-aggregating segments. Aromaticity, β-propensity, and formal charges play a major role in our model, as they are known in the literature to be determinant for fibrillization (Gazit 2002; Tjernberg et al. 2002; Chiti et al. 2003). Polar and nonpolar surfaces, as well as solubility, are also taken into account following an analysis of sequences designed to aggregate into β-sheets. The design of β-aggregating sequences was performed by structural sampling using molecular dynamics and peptide sequence optimization by a genetic algorithm (Tartaglia et al. 2004; G.G. Tartaglia and A. Caflisch, in prep.) (see subsection Derivation of the Equation). The aggregation propensity π_{il} of an l-residue segment starting at position i in the sequence is evaluated as:

$$\pi_{il} = \phi_{il}\Phi_{il} \tag{1}$$

The factor Φ_{il} contains exponential functions and is position-dependent

$$\Phi_{il} = e^{A_{il}+B_{il}+C_{il}} \tag{2}$$

where A_{il}, B_{il}, and C_{il} are functionals related to the aromaticity, β-propensity, and charge, respectively. The factor ϕ_{il} depends almost exclusively on the amino acid composition

$$\phi_{il} = \left[\prod_{j=i}^{i+l-1}\left(\frac{S_j^a}{\hat{S}^a}\theta^{\uparrow\uparrow} + \frac{S_j^p}{\hat{S}^p}\theta^{\uparrow\downarrow}\right)\frac{\hat{S}^t}{S_j^t}\frac{\hat{\sigma}}{\sigma_j}\right]^{1/l} \tag{3}$$

where S_i^a, S_i^p, S_i^t, and σ_i—weighted by their average over the 20 standard amino acids (hatted values)—are the side-chain apolar, polar, total water-accessible surface area, and solubility, respectively (see subsection Parallel and Anti-Parallel Configuration). The functionals $\theta^{\uparrow\uparrow}$ and $\theta^{\uparrow\downarrow}$ include positional effects and reflect the parallel or anti-parallel tendency to aggregate if the majority of residues is apolar or polar, respectively. Considering the high correlation between measured and predicted changes in aggregation rate upon single point mutations (Chiti et al. 2003; DuBay et al. 2004; Tartaglia et al. 2004), it is possible to utilize the propensity π_{il} to predict the absolute rate v_{il}

$$\nu_{il} = \alpha(c,T)\ \pi_{il} \tag{4}$$

where $\alpha(c,T)$ is introduced to take into account concentration and temperature (see subsection Concentration and Temperature).

Parallel and anti-parallel configuration

The functional for the parallel or anti-parallel configuration was introduced following the analysis of sequences designed by genetic algorithm optimization (see subsection Derivation of the Equation; Fig. 6):

- The parallel in-register β-sheet organization within fibrils is favored by the number of side chains involved in π-stacking (Tyr, Phe, and Trp) and apolar interactions (Ala, Gly, Ile, Leu, Met, Pro, and Val) (McGaughey et al. 1998; Azriel and Gazit 2001; Jenkins and Pickersgill 2001; Makin et al. 2005). The number of aromatic and apolar residues is indicated with $n_{aromatic}$ and n_{apolar}, respectively. Hydrogen bonds

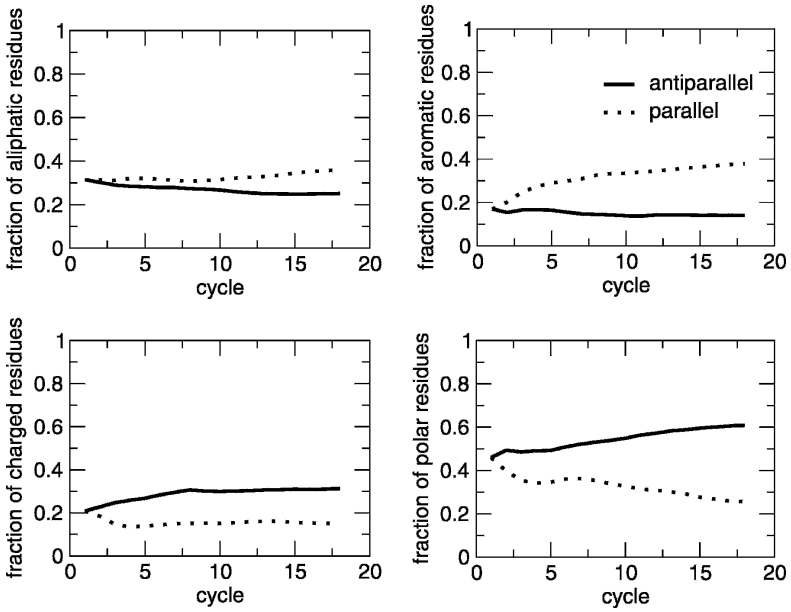

Figure 6. Computational design: A genetic algorithm approach was developed to search the space of peptide sequences for those with the best match to a given three-dimensional target conformation, i.e., an in-register parallel or anti-parallel aggregate of three heptapeptides (Gsponer et al. 2003). For each peptide sequence, three replicas were submitted to a 330 K molecular dynamics simulation, starting from the β-aggregated conformation using CHARMM parameter 19 and a solvent-accessible surface-based solvation model (Brooks et al. 1983; Ferrara et al. 2002). The sequence optimization was performed by evolutionary cycles. A total of 1728 sequences was sampled after 18 cycles. In sequences selected for the parallel aggregation, the number of aliphatic and aromatic residues increases almost monotonically, while the number of charged and polar residues decreases. The opposite is observed in sequences selected for the anti-parallel aggregation. In the plots, the number of aliphatic, aromatic, charged, and polar residues is normalized by the length of the peptide and averaged over the population (48 peptides per cycle).

between polar residues are not considered for the parallel aggregation because the number of polar residues decreases significantly during the optimization of parallel aggregated sequences (Fig. 6).

- The anti-parallel configuration is mainly determined by the electric dipole moment of the polypeptide (Hwang et al. 2004). Sequences abounding in polar residues show a small tendency for the parallel in-register aggregation because of unfavorable dipole–dipole interactions between side chains. Hence, the anti-parallel organization is promoted by the number of polar residues (Arg, Asn, Asp, Cys, Gln, Glu, His, Lys, Ser, and Thr), which is indicated with n_{polar}. In some specific positions, charged (Arg, Lys, Asp, and Glu) and aromatic amino acids contribute to the anti-parallel aggregation. "Specific positions" means that one or more couples of opposite charged residues or one or more aromatic residues are symmetrically placed with respect to the center of the sequence (Balbach et al. 2000; Hwang et al. 2004; Makin et al. 2005). In this specific case, the number of charged and aromatic residues is labeled as n^s_{charge} and $n^s_{aromatic}$, respectively.

In Equation 3, a parallel configuration is preferred if $n_{apolar} + n_{aromatic} > n_{polar} + n^s_{charge} + n^s_{aromatic}$. Since the number of aromatic residues in symmetric position is always smaller than the total amount of aromatic residues, $n_{aromatic} \geq n^s_{aromatic}$ (e.g., in the APP stretch: LVFFA $n_{aromatic} = 2$, $n^s_{aromatic} = 1$), we used a stricter condition for the parallel arrangement $n_{apolar} > n_{polar} + n^s_{charge}$. The stricter condition allows the factorization of aromatic contributions in Equation 1. In the ϕ_{il} factor of Equation 3, $\theta^{\uparrow\uparrow}$ and $\theta^{\uparrow\downarrow}$ are

$$\theta^{\uparrow\uparrow} = \begin{cases} 1 & n_{apolar} \geq n_{polar} + n^s_{charge} \\ 0 & otherwise \end{cases}$$
$$\theta^{\uparrow\downarrow} = 1 - \theta^{\uparrow\uparrow}$$

It is useful to explain the effect of the $\theta^{\uparrow\uparrow}$ and $\theta^{\uparrow\downarrow}$ functional by some examples. The segment LVFFA at position 671–676 of the APP is predicted to be parallel because it satisfies the parallel condition $n_{apolar} > n_{polar} + n^s_{charge}$ with $n_{apolar} = 3$ and $n_{polar} = n^s_{charge} = 0$ ($\theta^{\uparrow\uparrow} = 1$). The segment KLVFFAE (at position 670–677 of the APP), with two opposite charged residues, has anti-parallel propensity because it satisfies the anti-parallel condition $n_{apolar} < n_{polar} + n^s_{charge}$ with $n_{apolar} = 3$ and $n_{polar} = n^s_{charge} = 2$ ($\theta^{\uparrow\downarrow} = 1$).

The IAPP stretch FGAIL at position 22–26 is predicted to be parallel ($n_{apolar} = 4$ and $n_{polar} = n^s_{charge} = 0$, i.e.,$\theta^{\uparrow\uparrow} = 1$), in agreement with experimental results (Kayed et al. 1999a; Azriel and Gazit 2001; Gazit 2002). As in Azriel and Gazit (2001), the following stretches are predicted to be parallel: SVQFV at position 289–292 of lactadherin; DCFIL, CFILD,

and FILDL at position 187–191, 188–192, and 189–193 of gelsolin, respectively; FFSFL, FSFLG, and SFLGE at position 3–7, 4–8, and 5–9 of serum amyloid, respectively.

Poly(Gln), poly(Asn), and poly(Lys) homopolymers are predicted to be in an anti-parallel arrangement, as proposed in Perutz et al. (1994), Scherzinger et al. (1997), and Michelitsch and Weissman (2000) and observed by Tanaka et al. (2001) and Dzwolak et al. (2004). Moreover, it is likely that completely aliphatic sequences result in amorphous aggregates if N and C termini are capped, while a tendency to the anti-parallel arrangement is expected for short stretches with charged termini (e.g., transthyretin's stretch IAALL). Capping groups are neglected in the present version of the model.

The fragment GNNQQNY from the Sup35 yeast prion is predicted to be anti-parallel ($n_{apolar} = 1$, $n_{polar} = 5$, and $n^s_{charge} = 0$, i.e.,$\theta^{\uparrow\downarrow} = 1$), in contrast with the parallel packing suggested on the basis of X-ray diffraction and Fourier transform infrared (FTIR) data (Balbirnie et al. 2001). On one hand, it is important to note that the experimental data supporting a parallel arrangement are not conclusive, and, in particular, FTIR can be misleading on this point. In fact, in the unit cell of the microcrystals, the parallel β-sheets are proposed to be in anti-parallel contact along the fibril axis. On the other hand, a possible reason for the parallel configuration is that the π-interactions between the Tyr side chains are much less favorable in the anti-parallel configuration.

Aromatic residues

Aromatic side chains contribute to the parallel aggregation with π-interactions (McGaughey et al. 1998; Azriel and Gazit 2001; Makin et al. 2005). The density of aromatic residues $n_{aromatic}/l$ is used to distinguish two regimes for the aromatic contribution A_{il} of Equation 2:

$$A_{il} = \begin{cases} A^{low}_{il} & n_{aromatic}/l \le 3/20 \\ A^{high}_{il} & otherwise \end{cases}$$

where 3/20 is the aromatic density averaged over the 20 standard amino acids and $n_{aromatic}$ was defined in the previous subsection. In the case of low aromatic density ($n_{aromatic}/l \le 3/20$), A^{low}_{il} takes into account the polar/apolar environment. Following the results obtained by the genetic algorithm optimization of β-aggregation-prone sequences (see Fig. 6), A^{low}_{il} has a positive effect for mainly apolar sequences and a negative contribution for mainly polar sequences:

$$A^{low}_{il} = n_{aromatic}\,[n_{apolar} - (n_{polar} + n^s_{charge})]\,l^{-1}$$

The variables n_{apolar}, n_{polar}, and n^s_{charge} are defined in the previous subsection.

As an example, the APP stretch LVFFAEDVGSNKGAIIGLMVGGVVI shows low aromatic density ($n_{aromatic}/l = 2/25 < 3/20$). Since $i = 671$, $l = 25$, $n_{apolar} = 17$, $n_{polar} = 6$, and $n^s_{charge} = 0$, the aromatic contribution for LVFFAEDVGSNKGAIIGLMVGGVVI is $A^{low}_{671\ 25} = 2\ [17 - 6]\ 25^{-1} = 0.88$.

In the case of a high aromatic density ($n_{aromatic}/l > 3/20$), the model takes into account the number of aromatic residues:

$$A^{high}_{il} = n_{aromatic}$$

As an example, the APP stretch LVFFA shows high aromatic density ($n_{aromatic}/l = 2/5 > 3/20$). Since $i = 671$ and $l = 5$, the aromatic contribution for LVFFA is $A^{high}_{671\ 5} = 2$.

Besides the total amount of aromatic residues and the position dependence, which enters Equation 2 through A^{low}_{il}, the different polar and apolar side-chain surfaces, solubility, and β-propensity of Phe, Tyr, and Trp are taken into account in the factor ϕ_{il}. Hence, the mutation F22Y for the IAPP (islet β-amyloid protein precursor) stretch NFGAILSS produces a sensible change of rate ($\ln \pi_{wt} = -6$,$\ln \pi_{F22Y} = -7$), compatible with experiments in vitro (Porat et al. 2003).

β-Propensity

The β-propensity is evaluated as the fraction of residues that stabilize the β-sheet more than the α-helix:

$$B_{il} = \beta_{il}\, l^{-1} - 1/2$$

The function β_{il} is defined as:

$$\beta_{il} = \sum_{j=i}^{i+l-1} \delta^{\beta}_{j}$$

where

$$\delta^{\beta}_{j} = \begin{cases} 1 & \beta_j \ge \alpha_j \\ 0 & otherwise \end{cases}$$

The variables α_j and β_j correspond to the α-helix and β-sheet stabilizing effects of the amino acid at position j (Fersht 1999). Values of α_j and β_j are normalized from 0 (low stabilization) to 1 (high stabilization) to have the same range of variability. In the function B_{il}, the offset value of 1/2 is introduced so that $B_{il} > 0$ if at least one-half of the residues in the sequence is more stable in a β-sheet rather than in an α-helix conformation (i.e., $\beta_{il} > l^{-1}/2$).

In the case of the APP stretch LVFFA, values are $i = 671$, $l = 5$, $\beta_{672} = \beta_{673} = \beta_{674} = 1$, and $\beta_{671} = \beta_{675} = 0$. The predicted β-propensity for LVFFA is $\beta_{671\ 5} = 3/5 - 1/2 = 0.1$.

Charged residues

As in other models, we consider that the electrostatic repulsion of charged sequences penalizes the aggregation (Chiti et al. 2003; Tartaglia et al. 2004). In addition, our model takes into account the fact that side-chain pairs with opposite charges and positioned symmetrically with respect to the center of the segment contribute to the anti-parallel aggregation, as found in experiments (Gordon et al. 2004). In Equation 2, the charge contribution C_{il} is

$$C_{il} = -\frac{n_{charge}}{l}\left|\sum_{j=i}^{i+l-1} C_j\right| + \sum_{j=i}^{i+l-1} \delta^{charge}_{j}$$

where C_j is the charge of the side chain and n_{charge} is the number of charged residues. The first term of the functional C_{il} takes into account the electrostatic repulsion between polypeptides with net charge different from zero. The second term

counts the number of pairs of opposite charged side chains that are symmetrically placed with respect to the central residue of the sequence:

$$\delta_j^{charge} = \begin{cases} 1 & C_j = -C_{2i+l-j-1} \text{ and } C_j \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

In the case of the APP stretch KLVFFAE, the residues K_{670} and E_{676} have opposite charges and are symmetrically placed with respect to the central amino acid F_{673}. Since $i = 670$, $l = 7$, $C_{670} = +1$, and $C_{1340+7-670-1} = C_{676} = -1$, the net charge for KLVFFAE is $|\sum_{j=i}^{i+l-1} C_j| = 0$ and the oppositely charged K_{670} and E_{676} give $C_{670\ 7} = \delta_{670}^{charge} + \delta_{676}^{charge} = 2$.

Surfaces and solubility

For sequences that are predominantly apolar ($\theta^{\uparrow\uparrow} = 1$; see subsection Parallel and Anti-Parallel Configuration), the apolar water-accessible surface S_j^a measures the contribution of hydrophobic side chains to aggregation. For mostly polar sequences ($\theta^{\uparrow\downarrow} = 1$), the polar water-accessible surface S_j^p takes into account the propensity to form hydrogen bonds between polar residues. The total surface $S_j^t = S_j^a + S_j^p$ is used to weight polar and apolar surfaces by the total area. Values of apolar and polar side-chain surfaces are given in our previous work (Tartaglia et al. 2004) and span the intervals 44–195 $Å^2$ and 27–107 $Å^2$, respectively. Averaged values are $\hat{S}^a = 108\ Å^2$ and $\hat{S}^p = 54\ Å^2$. In the case of poly(Gln), values of surfaces are $S^a = 53\ Å^2$ and $S^p = 91\ Å^2$. Since Gln is polar and $\theta^{\uparrow\downarrow} = 1$, the surface contribution is $S^p/\hat{S}^p \cdot \hat{S}^t/S^t = (91/54)(162/144) = 1.9$.

The variable σ_j takes into account the water solubility of the side chain at position j. In our model, aggregation propensity and solubility are inversely proportional to introduce a penalty for highly soluble polypeptides. Most of the solubility values are available at http://acrux.igh.cnrs.fr/proteomics/densities_pi.html (Nahway 1989). The missing values (Cys, Lys, and Thr) were taken from http://www.formedium.com/Europe/amino_acids_and_vitamins.htm. The variable σ_j spans the interval 0.04–162 g/100 g, with average $\hat{\sigma} = 3.95$ g/100 g. In the case of poly(Gln), $\hat{\sigma}/\sigma = 3.95/2.5 = 1.5$, which indicates low solubility in agreement with experiments of β-aggregation (Perutz et al. 1994; Perutz 1999).

Concentration and temperature

The function $\alpha(c,T)$ captures the effects of concentration (c) and temperature (T) in Equation 4:

$$\alpha(c,T) = RT \begin{cases} c & c \in [0, c^*]\ mM \\ 1 & c \in (c^*, 1]\ mM \\ 1/c & c > 1\ mM \end{cases}$$

The aggregation rate v is approximated to be proportional to the temperature because the probability of collision and elongation of peptides increases with temperature (Kusumoto et al. 1998). Although aggregation rate and temperature are not expected to correlate above physiological values (Massi and Straub 2001), we used a simple linear dependence, which is preferable for the small extent of experimentally accessible values of the temperature. In fact, the temperature ranges from 298 K to 310 K in the data set of Figure 1.

In agreement with quasielastic light-scattering experiments of fibrillogenesis of $A\beta_{40}$, the aggregation rate v is assumed to be proportional to the concentration for $c < c^*$ mM ($c^* = 0.1$ mM) and to be independent of concentration above the critical value $c = c^*$ (Lomakin et al. 1996, 1997) (see also subsection-Derivation of the Equation). The hyperbolic function $1/c$ was introduced to decrease the aggregation rate v for $c > 1$ mM, as there is experimental evidence that a very high concentration opposes formation of ordered aggregates (Munishkina et al. 2004). The concentration ranges from 0.01 mM to 20 mM in the data set of Figure 1.

Derivation of the equation

- Functionals for aromaticity, β-propensity, and charge were taken from our relative rate equation (Tartaglia et al. 2004). The aromatic term was modified according to the results obtained by the genetic algorithm optimization of aggregating sequences (Fig. 6) (G.G. Tartaglia and A. Caflisch, in prep.). The functional for β-propensity, previously based on a single scale (Tartaglia et al. 2004), now takes into account β- versus α-propensity. Scales for β- and α-propensity are taken from Fersht (1999) and normalized in the range 0–1. The term used for the β-propensity was tested on 100 globular proteins: 82% of the β-sheet content is successfully recognized (data not shown). The functional for charged residues was modified with the addition of a term for symmetrically placed charges of opposite signs, which is consistent with experimental data (Gordon et al. 2004). The function n_{charge}/l replaces the constant factor in the relative rate (Tartaglia et al. 2004) and is introduced to weight the overall charge by the charge density. The three functionals for aromaticity, charge, and β-propensity can be zero. Exponential functions were introduced so that their product is different from zero.
- The product of the three functionals was plotted versus available experimental rates (see next subsection), obtaining a correlation of 80%, while the individual correlations for aromaticity, charge, and β-propensity are 76%, 81%, and 70%, respectively.
- The dependence on concentration and temperature was introduced to derive aggregation rates from propensities (Lomakin et al. 1997; Kusumoto et al. 1998; Massi and Straub 2001; Munishkina et al. 2004). With the concentration alone, the correlation improves to 85%. The correlation is 82% without the hyperbolic function for high concentrations ($c > 1$ mM). With the temperature function, the correlation improves to 88%.
- The factor for polar/apolar contributions ϕ_{il} in Equation 1 was added upon the analysis of sequences produced by computational design (Fig. 6). The term is a linear combination of normalized surfaces and has a nonzero minimum. The correlation improves to 92%. The solubility dependence was added at the very end and introduces a penalty for highly soluble sequences. The correlation improves to 95%.

Experimental data

Most of the experimental rates were kindly provided by Dr. F. Chiti and Dr. M. Vendruscolo (Chiti et al. 2003; DuBay et al. 2004). The remainder data set was taken from previous experimental studies (Litvinovich et al. 1998; Konno et al. 1999; Ferguson et al. 2003). The absolute aggregation rates were

determined from in vitro experiments of denaturated polypeptide chains without taking into account the presence of cellular components as chaperones and proteases. Aggregation rates were obtained from kinetic traces in different ways: thioflavin T fluorescence, turbidity, CD, sedimentation, size exclusion chromatography, and filtration. Lag phases were not considered in the analysis, as they were not reported or difficult to extract from published data (DuBay et al. 2004). Since a comprehensive understanding of lag phases in protein aggregation is lacking (Kayed et al. 1999b; Padrick and Miranker 2002) (e.g., it is not known whether fibrils form by addition of monomers or oligomers and how growth conditions influence the amyloid formation), the aggregation kinetics was analyzed after the lag phase. The elongation phase showing an exponential behavior is fitted to the function $z = \alpha (1 - e^{-vt})$ where v is the rate measured in $\sec^{-1}$.

Acknowledgments

We thank Prof. C. Dobson, Prof. F. Chiti, Dr. M. Vendruscolo, and Dr. J. Zurdo for providing rates of several proteins. The molecular dynamics simulations were performed on the Matterhorn Beowulf cluster at the Informatikdienste at the University of Zurich. We thank C. Bolliger, Dr. T. Steenbock, and Dr. A. Godknecht for setting up and maintaining the cluster. This work was supported by the Swiss National Science Foundation and the NCCR "Neural Plasticity and Repair."

The program for the calculation of aggregation rates is available from the corresponding author upon request.

References

Antzutkin, O.N., Balbach, J.J., Leapman, R.D., Rizzo, N.W., Reed, J., and Tycko, R. 2000. Multiple quantum solid-state NMR indicates a parallel, not antiparallel, organization of β-sheets in Alzheimer's β-amyloid fibrils. *Proc. Natl. Acad. Sci.* **97:** 13045–13050.

Antzutkin, O.N., Leapman, R.D., Balbach, J.J., and Tycko, R. 2002. Supramolecular structural constraints on Alzheimer's-amyloid fibrils from electron microscopy and solid-state nuclear magnetic resonance. *Biochemistry* **41:** 15436–15450.

Asl, L.H., Liepnieks, J.J., Uemichi, T., Rebibou, J.M., Justrabo, E., Droz, D., Mousson, J.M.C., Benson, M.D., Delpech, M., and Grateau, G. 1997. Renal amyloidosis with a frame shift mutation in fibrinogen α-chain gene producing a novel amyloid protein. *Blood* **90:** 4799–4805.

Azriel, R. and Gazit, E. 2001. Analysis of the minimal amyloid-forming fragment of the Islet amyloid polypeptide. *J. Biol. Chem.* **276:** 34156–34161.

Balbach, J.J., Ishii, Y., Antzutkin, O.N., Leapman, R.D., Rizzo, N.W., Dyda, F., Reed, J., and Tycko, R. 2000. Amyloid fibril formation by a β16–22, a seven-residue fragment of the Alzheimer's β-amyloid peptide, and structural characterization by solid state NMR. *Biochemistry* **39:** 13748–13759.

Balbirnie, M., Grothe, R., and Eisenberg, D. 2001. An amyloid-forming peptide from the yeast prion Sup35 reveals a dehydrated β-sheet structure for amyloid. *Proc. Natl. Acad. Sci.* **98:** 2375–2380.

Bitan, G., Kirkitadze, M.D., Lomakin, A., Vollers, S.S., Benedek, G.B., and Teplow, B.D. 2003. Amyloid Aβ-protein Aβ assembly: Aβ40 and Aβ42 oligomerize through distinct pathways. *Proc. Natl. Acad. Sci.* **100:** 330–335.

Bond, J.P., Deverin, S.P., Inouye, H., El-Agnaf, O.M.A, Teeter, M.M., and Kirschnera, D.A. 2003. Assemblies of Alzheimer's peptides Aβ25–35 and Aβ31–35: Reverse-turn conformation and side-chain interactions revealed by x-ray diffraction. *J. Struct. Biol.* **141:** 156–170.

Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S., and Karplus, M. 1983. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **4:** 187–217.

Broome, B.M. and Hecht, M.H. 2000. Nature disfavors sequences of alternating polar and non-polar amino acids: Implications for amyloidogenesis. *J. Mol. Biol.* **296:** 961–968.

Chiti, F., Calamai, M., Taddei, N., Stefani, M., Ramponi, G., and Dobson, C.M. 1999. Studies of the aggregation of mutant proteins in vitro provide insights into the genetics of amyloid diseases. *Proc. Natl. Acad. Sci.* **99:** 16419–16426.

Chiti, F., Stefani, M., Taddei, N., Ramponi, G., and Dobson, C.M. 2003. Rationalization of the effects of mutations on peptide and protein aggregation rates. *Nature* **424:** 805–808.

Dobson, C.M. 1999. Protein misfolding, evolution and disease. *Trends Biochem. Sci.* **24:** 329–332.

DuBay, K.F., Pawar, A.P., Chiti, F., Zurdo, J., Dobson, C.M., and Vendruscolo, M. 2004. Predicting absolute aggregation rates of amyloidogenic polypeptide chains. *J. Mol. Biol.* **341:** 1317–1326.

Dzwolak, W., Muraki, T., Kato, M., and Taniguchi, Y. 2004. Chain-length dependence of α-helix to β-sheet transition in polylysine: Model of protein aggregation studied by temperature-tuned FTIR spectroscopy. *Biopolymers* **73:** 463–469.

El-Agnaf, O.M.A., Sheridan, J.M., Sidera, C., Siligardi, G., Hussain, R., Haris, P.I., and Austen, B.M. 2001. Effect of the disulfide bridge and the C-terminal extension on the oligomerization of the amyloid peptide ABri implicated in familial British dementia. *Biochemistry* **40:** 3449–3457.

El-Agnaf, O.M.A., Gibson, G., Lee, M., Wright, A., and Austen, B.M. 2004. Properties of neurotoxic peptides related to the Bri gene. *Protein Pept. Lett.* **11:** 202–212.

Ferguson, N., Berriman, J., Petrovich, M., Sharpe, T.D., Finch, J.T., and Fersht, A.R. 2003. Rapid amyloid fibril formation from the fast-folding WW domain FBP28. *Proc. Natl. Acad. Sci.* **100:** 9814–9819.

Fernandez Escamilla, A.M., Rousseau, F., Schymkowitz, J., and Serrano, L. 2004. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat. Biotech.* **22:** 1302–1306.

Ferrara, P., Apostolakis, J., and Caflisch, A. 2002. Evaluation of a fast implicit solvent model for molecular dynamics simulations. *Proteins* **46:** 24–33.

Fersht, A.R. 1999. *Structure and mechanism in protein science*. Freeman and Co., New York.

Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R.D., and Bairoch, A. 2003. Expasy: The proteomics server for in depth protein knowledge and analysis. *Nucleic Acids Res.* **31:** 3784–3788.

Gazit, E. 2002. A possible role for π-stacking in the self-assembly of amyloid fibrils. *FASEB J.* **16:** 77–83.

Gordon, D.J., Balbach, J.J., Tycko, R., and Meredith, S.C. 2004. Increasing the amphiphilicity of an amyloidogenic peptide changes the β-sheet structure in the fibrils from antiparallel to parallel. *Biophys. J.* **86:** 428–434.

Gsponer, J., Habertuer, U., and Caflisch, A. 2003. The role of side-chain interactions in the early steps of aggregation: Molecular dynamics simulations of an amyloid-forming peptide from the yeast prion Sup35. *Proc. Natl. Acad. Sci.* **100:** 5154–5159.

Haggqvist, B., Naeslund, J., Sletten, K., Westermark, G.T., Mucchiano, G., Tjernberg, L.O., Nordstedt, C., Engstroem, U., and Westermark, P. 1999. Medin: An integral fragment of aortic smooth muscle cell-produced lactadherin forms the most common human amyloid. *Proc. Natl. Acad. Sci.* **96:** 8669–8674.

Hill, A.F., Joiner, S., Linehan, J., Desbruslais, M., Lantos, P.L., and Collinge, J. 2000. Species-barrier-independent prion replicates in apparently resistant species. *Proc. Natl. Acad. Sci.* **97:** 10248–10253.

Horwich, A.L. and Weissman, J.S. 1997. Deadly conformations-protein misfolding disease. *Cell* **89:** 499–510.

Hwang, W., Zhang, S., Kamm, R.D., and Karplus, M. 2004. Kinetic control of dimer structure formation in amyloid fibrillogenesis. *Proc. Natl. Acad. Sci.* **101:** 12916–12921.

Jaroniec, C.P., MacPhee, C.E., Astrof, N.S., Dobson, C.M., and Griffin, R.G. 2002. Molecular conformation of a peptide fragment of transthyretin in an amyloid fibril. *Proc. Natl. Acad. Sci.* **99:** 16748–16753.

Jarrett, J., Berger, E.P., and Lansbury Jr., P.T. 1993. The carboxyl terminus of the β amyloid protein critical for the seeding of amyloid formation: Implications for the pathogenesis of Alzheimer's disease. *Biochemistry* **32:** 4693–4697.

Jenkins, J. and Pickersgill, R. 2001. The architecture of parallel β-helices and related folds. *Prog. Biophys. Mol. Biol.* **77:** 111–115.

Jimenez, J.L., Nettleton, E.J., Bouchard, M., Robinson, C.V., Dobson, C.M., and Saibil, H.R. 2002. The protofilament structure of insulin amyloid fibrils. *Proc. Natl. Acad. Sci.* **99:** 9196–9201.

Jones, S., Manning, J., Kad, N.M., and Radford, S.E. 2003. Amyloid-forming peptides from b_2 microglobulin—Insights into the mechanism of fibril formation in vitro. *J. Mol. Biol.* **325:** 249–257.

Kangas, H., Paunio, T., Kalkkinen, N., Jalanko, A., and Peltonen, L. 1996. In vitro expression analysis shows that the secretory form of Gelsolin is

the sole source of amyloid in Gelsolin-related amyloidosis. *Hum. Mol. Genet.* **5:** 1237–1244.

Kayed, R., Bernhagen, J., Greenfield, N., Sweimeh, K., Brummer, H., Voelter, W., and Kapurniotu, A. 1999a. Conformational transitions of islet amyloid polypeptide (IAPP) in amyloid formation in vitro. *J. Mol. Biol.* **287:** 781–796.

———. 1999b. Partial molar volume, surface area, and hydration changes for equilibrium unfolding and formation of aggregation transition state: High-pressure and cosolute studies on recombinant human IFN-γ. *J. Mol. Biol.* **287:** 781–796.

Kelly, J. 1998. The alternative conformations of amyloidogenic proteins and their multi-step assembly pathways. *Curr. Opin. Struct. Biol.* **8:** 101–106.

King, C.Y., Tittmann, P., Gross, H., Gebert, R., Aebi, M., and Wuethrich, K. 1997. Prion-inducing domain 2–114 of yeast Sup35 protein transforms in vitro into amyloid-like filaments. *Proc. Natl. Acad. Sci.* **94:** 6618–6622.

Konno, T., Murata, K., and Nagayama, K. 1999. Amyloid-like aggregates of a plant protein: A case of sweet tasting protein, monellin. *FEBS Lett.* **454:** 122–126.

Kozin, S.A., Bertho, G., Mazur, A.K., Rabesona, H., Girault, J.P., Haerlthé, T., Takahashi, M., Debey, P., and Hui Bon Hoa, G. 2001. Sheep prion protein synthetic peptide spanning helix 1 and β-strand 2 residues 142–166 shows β-hairpin structure in solution. *J. Biol. Chem.* **49:** 46364–46370.

Kusumoto, Y., Lomakin, A., Teplow, D.B., and Benedek, G.B. 1998. Temperature dependence of amyloid β-protein fibrillization. *Proc. Natl. Acad. Sci.* **95:** 12277–12282.

Linding, R., Schymkowitz, J., Rousseau, J., Diella, F., and Serrano, L. 2004. A comparative study of the relationship between protein structure and β-aggregation in globular and intrinsically disordered proteins. *J. Mol. Biol.* **342:** 345–353.

Litvinovich, S.V., Brew, S.A., Aota, S., Akiyama, S.K., Haudenschild, C., and Ingham, K.C. 1998. Formation of amyloid like fibrils by self-association of a partially unfolded fibronectin type III module. *J. Mol. Biol.* **280:** 245–258.

Lomakin, A., Chung, D.S., Benedek, G.B., Kirschner, D.A., and Teplow, D.B. 1996. On the nucleation and growth of amyloid β-protein fibrils: Detection of nuclei and quantitation of rate constants. *Proc. Natl. Acad. Sci.* **93:** 1125–1129.

Lomakin, A., Teplow, D.B., Kirschner, D.A., and Benedek, G.B. 1997. Kinetic theory of fibrillogenesis of amyloid β-protein. *Proc. Natl. Acad. Sci.* **94:** 7942–7947.

Makin, O.S., Atkins, E., Sikorski, P., Johansson, J., and Serpell, L.C. 2005. Molecular basis for amyloid fibril formation and stability. *Proc. Natl. Acad. Sci.* **102:** 315–320.

Marcotte, E.M. and Eisenberg, D. 1999. Chicken prion tandem repeats form a stable, protease-resistant domain. *Biochemistry* **38:** 667–676.

Margittai, M. and Langen, R. 2004. Template-assisted filament growth by parallel stacking of τ. *Proc. Natl. Acad. Sci.* **101:** 10279–10283.

Massi, F. and Straub, J.E. 2001. Energy landscape theory for Alzheimer's amyloid β-peptide fibril elongation. *Proteins* **42:** 217–229.

Matthews, D. and Cooke, B. 2003. The potential for transmissible spongiform encephalopathies in non-ruminant livestock and fish. *Rev. Sci. Tech.* **22:** 283–296.

McGaughey, G.B., Gagné, M., and Rappé, A.K. 1998. π-Stacking interaction. *J. Biol. Chem.* **273:** 15458–15463.

Michelitsch, M.D. and Weissman, J.S. 2000. A census of glutamine/asparagine-rich regions: Implications for their conserved function and the prediction of novel prions. *Proc. Natl. Acad. Sci.* **97:** 11910–11915.

Munishkina, L.A., Fink, A.L., and Uversky, V.U. 2004. Conformational prerequisites for formation of amyloid fibrils from histones. *J. Mol. Biol.* **342:** 1305–1324.

Nahway, N. 1989. *The Merck index*. Merck and Co., Inc., Whitehouse Station, NJ.

Nguyen, J., Baldwin, M.A., Cohen, F.E., and Prusiner, S.B. 1995. Prion protein peptides induce α-helix to β-sheet conformational transitions. *Biochemistry* **34:** 4186–4192.

Nichols, W.C., Dwulet, F.E., Liepnieks, J., and Benson, M.D. 1988. Variant apolipoprotein AI as a major constituent of a human hereditary amyloid. *Biochem. Biophys. Res. Commun.* **156:** 762–768.

Padrick, S.B. and Miranker, A.D. 2002. Islet amyloid: Phase partitioning and secondary nucleation are central to the mechanism of fibrillogenesis. *Biochemistry* **41:** 4694–4703.

Perutz, M.F. 1999. Glutamine repeats and neurodegenerative diseases: Molecular aspects. *Trends Biochem. Sci.* **24:** 58–64.

Perutz, M.F., Johnson, T., Suzuki, M., and Finch, J.T. 1994. Glutamine repeats as polar zippers: Their possible role in inherited neurodegenerative diseases. *Proc. Natl. Acad. Sci.* **91:** 5355–5358.

Porat, Y., Stepensky, A., Ding, F.X., Naider, F., and Gazit, E. 2003. Completely different amyloidogenic potential of nearly identical peptide fragments. *Biopolymers* **69:** 161–163.

Prusiner, S.B. 1988. Prions. *Proc. Natl. Acad. Sci.* **95:** 13363–13383.

Rochet, J.C. and Lansbury Jr., P.T. 2000. Amyloid fibrillogenesis: Themes and variations. *Curr. Opin. Struct. Biol.* **10:** 60–68.

Scherzinger, E., Lurz, R., Turmaine, M., Mangiarini, L., Hollenback, B., Hasenbank, R., Bates, G.P., Davies, S.W., Lehrack, H., and Wanker, E. 1997. Huntingtin-encoded polyglutamine expansions form amyloid-like protein aggregates in vitro and in vivo. *Cell* **90:** 549–558.

Soto, C. and Castilla, J. 2004. The controversial protein-only hypothesis of prion propagation. *Nat. Med.* **10:** S63–S67.

Stefani, M. and Dobson, C.M. 2003. Protein aggregation and aggregate toxicity: New insights into protein folding, misfolding diseases and biological evolution. *J. Mol. Med.* **81:** 678–699.

Tanaka, M., Morishima, I., Akagi, T., Hashikawa, T., and Nukina, N. 2001. Intra and intermolecular β-pleated sheet formation in glutamine-repeat inserted myoglobin as a model for polyglutamine diseases. *J. Biol. Chem.* **276:** 45470–45475.

Tartaglia, G.G., Cavalli, A., Pellarin, R., and Caflisch, A. 2004. The role of aromaticity, exposed surface, and dipole moment in determining protein aggregation rates. *Protein Sci.* **13:** 1939–1941.

Tartaglia, G.G., Pellarin, R., Cavalli, A., and Caflisch, A. 2005. Organism complexity anti-correlates with proteomic β-aggregation propensity. *Protein Sci.* (this issue).

Tenidis, K., Waldner, M., Bernhagen, J., Fischle, W., Bermann, M., Weber, M., Merkle, M., Voelter, W., Brunner, H., and Kapurniotu, A. 2000. Identification of a penta- and hexapeptide of Islet amyloid polypeptide IAPP with amyloidogenic and cytotoxic properties. *J. Mol. Biol.* **295:** 1055–1071.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. 1994. Clustal W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22:** 4673–4680.

Tjernberg, L., Hosia, W., Bark, N., Thyberg, J., and Johansson, J. 2002. Charge attraction and β-propensity are necessary for amyloid fibril formation from tetrapeptides. *J. Biol. Chem.* **277:** 43243–43246.

Torok, M., Milton, S., Kayed, R., Wu, P., Intire, T.M., Glabe, C., and Langen, R. 2002. Structural and dynamic features of Alzheimer Aβ peptide in amyloid fibrils studied by site-directed spin labeling. *J. Biol. Chem.* **277:** 40810–40815.

Ueda, K., Fukushima, H., Masliah, E., Xia, Y., Iwai, A., Yoshimoto, M., Otero, D.A., Kondo, J., Ihara, Y., and Saitoh, T. 1993. Molecular cloning of cDNA encoding an unrecognized component of amyloid in Alzheimer disease. *Proc. Natl. Acad. Sci.* **90:** 11282–11286.

Vanik, D.L., Surewicz, K.A., and Surewicz, W.K. 2004. Molecular basis of barriers for intraspecies transmissibility of mammalian prions. *Mol. Cell* **14:** 139–145.

von Bergen, M., Friedhoff, P., Biernat, J., Heberle, J., Mandelkow, E.M., and Mandelkow, E. 2000. Assembly of τ protein into Alzheimer paired helical filaments depends on a local sequence motif (^{306}VQIVYK^{311}) forming β-structure. *Proc. Natl. Acad. Sci.* **97:** 5129–5134.

Weidemann, A., Konig, G., Bunke, D., Fisher, P., Salbaum, J.M., Masters, C.L., and Beyreuther, K. 1989. Identification, biogenesis and localization of precursors of Alzheimer's disease A4 amyloid protein. *Cell* **57:** 115–126.

Westermark, P., Wernstedt, C., Wilander, E., Hayden, D.W., O'Brien, T.D., and Johnson, K.H. 1987. Amyloid fibrils in human insulinoma and islets of Langerhans of the diabetic cat are derived from a neuropeptide-like protein also present in normal islet cells. *Proc. Natl. Acad. Sci.* **84:** 3881–3885.

Westermark, G.T., Engstrom, U., and Westermark, P. 1992. The N-terminal segment of protein AA determines its fibrillogenic propensity. *Biochem. Biophys. Res. Commun.* **182:** 27–32.

Williams, A.D., Portelius, E., Kheterpal, I., Guo, J.T., Cook, K.D., Xu, Y., and Wetzel, R. 2004. Mapping Aβ amyloid fibril secondary structure using scanning proline mutagenesis. *J. Mol. Biol.* **335:** 833–842.

Zoete, V., Michielin, O., and Karplus, M. 2003. Protein-ligand binding free energy estimation using molecular mechanics and continuum electrostatics. Application to HIV-1 protease inhibitors. *J. Comput. Aided Mol. Des.* **17:** 861–880.

Chapter 5

Organism Complexity Anticorrelates with Proteomic β- Aggregation Propensity

Protein Science (2005) 14, 2735-2740

FOR THE RECORD

Organism complexity anti-correlates with proteomic β-aggregation propensity

GIAN GAETANO TARTAGLIA,[1] RICCARDO PELLARIN,[1] ANDREA CAVALLI, AND AMEDEO CAFLISCH

Department of Biochemistry, University of Zürich, CH-8057 Zürich, Switzerland

(RECEIVED March 23, 2005; FINAL REVISION June 23, 2005; ACCEPTED June 24, 2005)

Abstract

We introduce a novel approach to estimate differences in the β-aggregation potential of eukaryotic proteomes. The approach is based on a statistical analysis of the β-aggregation propensity of polypeptide segments, which is calculated by an equation derived from first principles using the physicochemical properties of the natural amino acids. Our analysis reveals a significant decreasing trend of the overall β-aggregation tendency with increasing organism complexity and longevity. A comparison with randomized proteomes shows that natural proteomes have a higher degree of polarization in both low and high β-aggregation prone sequences. The former originates from the requirement of intrinsically disordered proteins, whereas the latter originates from the necessity of proteins with a stable folded structure.

Keywords: aggregation; protein aggregation propensity; proteome; intrinsically disordered proteins

Supplemental material: see www.proteinscience.org

Even proteins not implicated in amyloid diseases have been shown to form fibrils in vitro under denaturing conditions, indicating that fibrillogenesis is a common feature of polypeptide chains, which can form intermolecular backbone–backbone hydrogen bonds (Chiti et al. 1999, 2003) and favorable side-chain interactions (Azriel and Gazit 2001; Gsponer et al. 2003; Makin et al. 2005). Although in lower eukaryotes amyloid fibrils could represent an inheritable phenotype related to specific cellular functions (Osherovich and Weissman 2002; Osherovich et al. 2004; Si et al. 2003b), the cytotoxicity of prefibrillar aggregates (Bucciantini et al. 2002) and their association with diseases such as Alzheimer's, Parkinson's, Huntington's, prion disease, cystic fibrosis, and type II diabetes (Kelly 1998; Rochet and Lansbury 2000) suggest that amyloid aggregates are generally dangerous for higher eukaryotes (Dobson 1999; Stefani and Dobson 2003).

We have previously developed an equation to predict the propensity for ordered aggregation, which solely requires the polypeptide sequence as input (Tartaglia et al. 2004, 2005). Our model is based on the physicochemical properties of the residues and takes into account both amino acid composition and positional information. The aggregation propensity π_{il} of an l-residue segment starting at position i in the sequence is evaluated as

$$\pi_{il} = \phi_{il}\ \Phi_{il} \tag{1}$$

The factor Φ_{il} contains exponential functions and is position-dependent

$$\Phi_{il} = e^{A_{il}+B_{il}+C_{il}} \tag{2}$$

where A_{il}, B_{il}, and C_{il} are functionals related to the aromaticity, β-propensity, and charge, respectively. The fac-

[1]These authors contributed equally to this work.

Reprint requests to: Gian Gaetano Tartaglia or Amedeo Caflisch, Department of Biochemistry, University of Zürich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland; e-mail: gian@bioc.unizh.ch or caflisch@bioc.unizh.ch; fax: +41-44-635-68-62.

Article published online ahead of print. Article and publication date are at http://www.proteinscience.org/cgi/doi/10.1110/ps.051473805.

tor ϕ_{il} depends almost exclusively on the amino acid composition

$$\phi_{il} = \left[\prod_{j=i}^{i+l-1}\left(\frac{S_j^a}{\hat{S}^a}\theta^{\uparrow\uparrow} + \frac{S_j^p}{\hat{S}^p}\theta^{\uparrow\downarrow}\right)\frac{\hat{S}^t}{S_j^t}\frac{\hat{\sigma}}{\sigma_j}\right]^{1/l} \quad (3)$$

where S_i^a, S_i^p, S_i^t, and σ_i—weighted by their average over the 20 standard amino acids (hatted values)—are the side-chain apolar, polar, total water-accessible surface area, and solubility, respectively. The functionals $\theta^{\uparrow\uparrow}$ and $\theta^{\uparrow\downarrow}$ include positional effects and reflect the parallel or antiparallel tendency to aggregate if the majority of residues is apolar or polar, respectively. Details of the method are presented in the preceding paper (Tartaglia et al. 2005).

In the present work, we analyze complete proteomes of several eukaryotes to identify changes of β-aggregation propensity through organisms of different complexity. The 32,869 entries belonging to the human proteome database (Supplemental Material, Table 1) were decomposed in stretches of different sizes (5, 50, and 150 residues) to compute the β-aggregation propensity with Equation 1 and build the normalized histogram of β-aggregation propensity distribution, APD (Fig. 1A). For each stretch size, the distribution is found to be nonsymmetric with respect to the average and skewed to the left, indicating that there are more stretches with low β-aggregation propensity (left tail of APD) than with high propensity (right tail). As pointed out in our previous study, short stretches are preferable to long stretches for the analysis of β-aggregation propensity because the latter contain folding features that deteriorate the signal-to-noise ratio (Tartaglia et al. 2005).

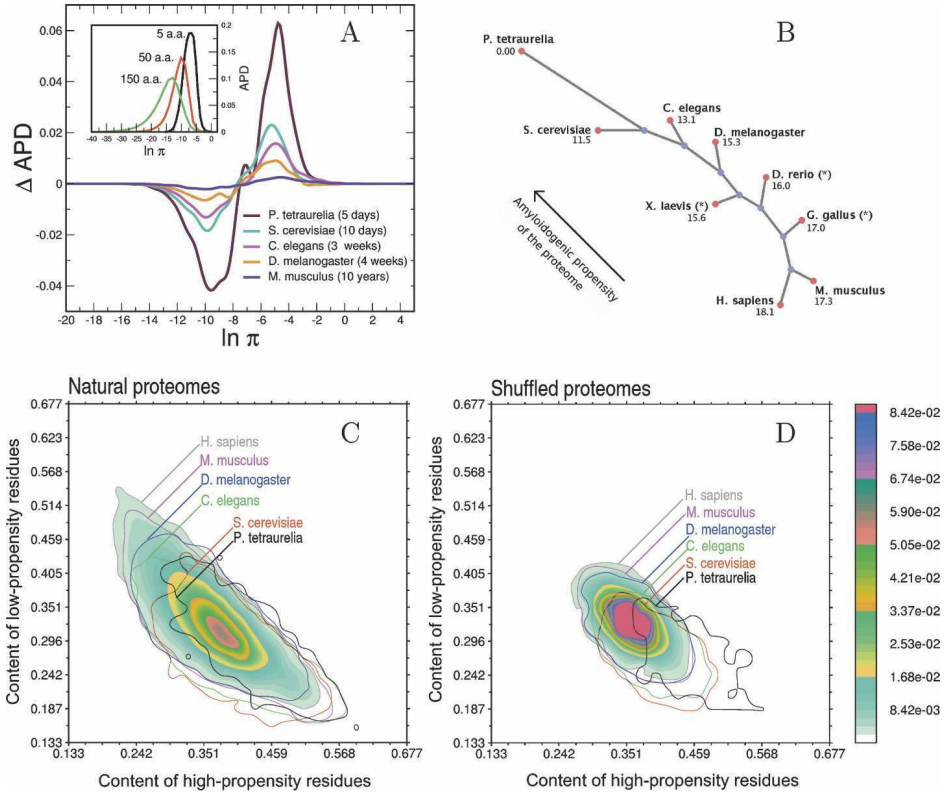

Figure 1. (*A*) (*Inset*) Distribution of the number of human polypeptide sequences as a function of β-aggregation propensity (APD) at three different window sizes. (*Main plot*) APD differences with respect to *H. sapiens* for complete proteomes of *M. musculus*, *D. melanogaster*, *C. elegans*, *S. cerevisiae*, and *P. tetraurelia* (window size of five residues). Life spans of organisms are reported in parentheses. (*B*) Unrooted tree diagram derived from the APD deviation (Equation 4). The deviation is computed from *P. tetraurelia* as a reference and magnified by a factor of 1000. The arrow indicates that lower eukaryotes have more high-propensity and fewer low-propensity stretches. This diagram is built using Phylodraw with the Fitch and Margoliash (1967) clustering algorithm. Data labeled with * belong to incomplete proteomes. (Phylodraw is available at http://pearl.cs.pusan.ac.kr/phylodraw/.) (*C*) Normalized histogram of the number of proteins as a function of the content of residues enriched in low-propensity and high-propensity stretches. Global contours are shown for all proteomes by solid lines. Isofrequency regions are shown for the human proteome, where red color indicates the most populated area, while blue fading color indicates the least-populated areas. (*D*) Same as *C* for shuffled proteomes.

Hence, a window size of five residues was used to analyze complete proteomes of *Homo sapiens*, *Mus musculus*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae*, and *Paramecium tetraurelia* (Supplemental Material, Table 1). Nonhuman eukaryotes show a larger amount of high-propensity stretches and a smaller amount of low-propensity stretches compared with *H. sapiens* (Fig. 1A). Moreover, a clear trend is found with the increasing complexity of the organisms and their lifetime. To quantify this trend it is useful to introduce the APD deviation between two proteomes, x and y

$$d_{xy} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(APD_x(\pi_i) - APD_y(\pi_i)\right)^2} \qquad (4)$$

where the β-aggregation propensity π is calculated by Equation 1 (Tartaglia et al. 2005) and i runs over the total number of bins N ($N = 100$) in the APD histogram. With the addition of the proteomes of *Danio rerio*, *Xenopus laevis*, and *Gallus gallus*, the APD deviation was used to build the tree diagram of Figure 1B. Except for the inversion between the amphibious *X. laevis* and the fish *D. rerio* (whose proteomes are not complete), the tree of Figure 1B is similar to the phylogenetic tree of cytochrome c (Dayhoff et al. 1972). Thus, the deviation calculated from *P. tetraurelia*, d_{xP}, is an observable able to rank proteomes of organisms of increasing complexity. It is interesting to compare the amino acid frequencies in APD tails—defined for a subtended area of 0.05 in the histogram of Figure 1A—with amino acid frequencies in entire proteomes (Table 1). This analysis reveals that for all proteomes stretches with low β-aggregation propensity are rich in *A*, *G*, *H*, *K*, *P* and *R*, whereas high-propensity stretches in *C*, *F*, *I*, *L*, *N*, *Q*, *V*, and *Y*. Figure 1C is a two-dimensional histogram that shows the number of proteins as a function of the content of residues enriched in low-propensity stretches and the content of residues predominant in high-propensity stretches. By increasing the organism complexity, the number of proteins with low-propensity residues increases, while the number of proteins with high-propensity residues decreases. A comparison with randomized proteomes is useful to further investigate the significance of such trends. Randomized proteomes were generated by shuffling amino acids within complete proteomes and keeping unchanged the global amino acid composition, number, and length of proteins. We stress that the β-aggregation propensity of five-residue stretches cannot differentiate natural and shuffled proteomes, because short segments describe mainly effects of the amino acid composition. Yet, differences between natural and shuffled proteomes are enhanced when residues belonging to low-/high-propensity stretches are used for the analysis of entire proteins. Comparing Figure 1, C and D, it is evident that shuffled proteomes are less spread. In other words, natural proteomes reveal a sensible increase of sequences with residues predominant in low-propensity stretches as well as residues enriched in high-propensity stretches. While the amino acid global composition of proteomes is almost identical in higher eukaryotes, the content of low-propensity stretches increases significantly, indicating a clear change of protein features from proteome to proteome.

Table 1. *Amino acid frequencies in left or right* APD *tails of* H. sapiens *divided by their corresponding frequency in the whole proteome*

	A	C	D	E	F	G	H	I	K	L
Left/total	**1.1**	0.2	0.4	0.5	0.4	**1.3**	**1.6**	0.5	**2.1**	0.5
Right/total	0.7	**2.4**	0.8	0.7	**2.7**	0.6	0.5	**1.6**	0.3	**1.8**
	M	N	P	Q	R	S	T	V	W	Y
Left/total	0.4	0.2	**3.3**	0.3	**2.8**	0.6	0.5	0.7	0.4	0.1
Right/total	0.8	**1.5**	0.2	**1.2**	0.2	0.7	0.8	**1.2**	0.8	**2.7**

Values exceeding 1.0 are shown in bold. Similar frequencies were found for all the proteomes.

It has recently been shown that natively unfolded proteins (or intrinsically disordered proteins, IDPs) are implicated in cellular regulation, signaling, and assembly/disassembly of macromolecular complexes (Dunker et al. 2002; Ward et al. 2004; Oldfield et al. 2005). The absence of a fixed structure suggests functional implications, which are required in complex organisms (Koonin et al. 2002). Interestingly, a larger diffusion of IDPs is found in higher eukaryotes than in lower eukaryotes and prokaryotes (Dunker et al. 2002; Liu et al. 2002; Linding et al. 2004). Using data from X-ray crystallography, nuclear magnetic resonance, and circular dichroism, Williams et al. (2001) found a high percentage of *P*, *R*, *K*, *G*, *A*, *Q*, *S*, and *E* in nonfolded segments of proteins, and *F*, *Y*, *C*, *L*, *V*, *N*, and *W* in folded segments. Except for *Q*, *S*, and *E*, Williams' finding is in agreement with our tail composition analysis (Table 1), indicating that residues enriched in aggregating stretches promote both folding and β-aggregation, whereas residues predominant in stretches with low β-aggregation propensity are also enriched in IDPs.

To better understand the relationship between β-aggregation propensity and protein structure, we analyzed the APDs of polypeptide segments that assume a regular secondary structure, as well as IDPs (Supplemental Material, Table 1). As shown in Figure 2A, strands have more β-aggregation potential than helices,

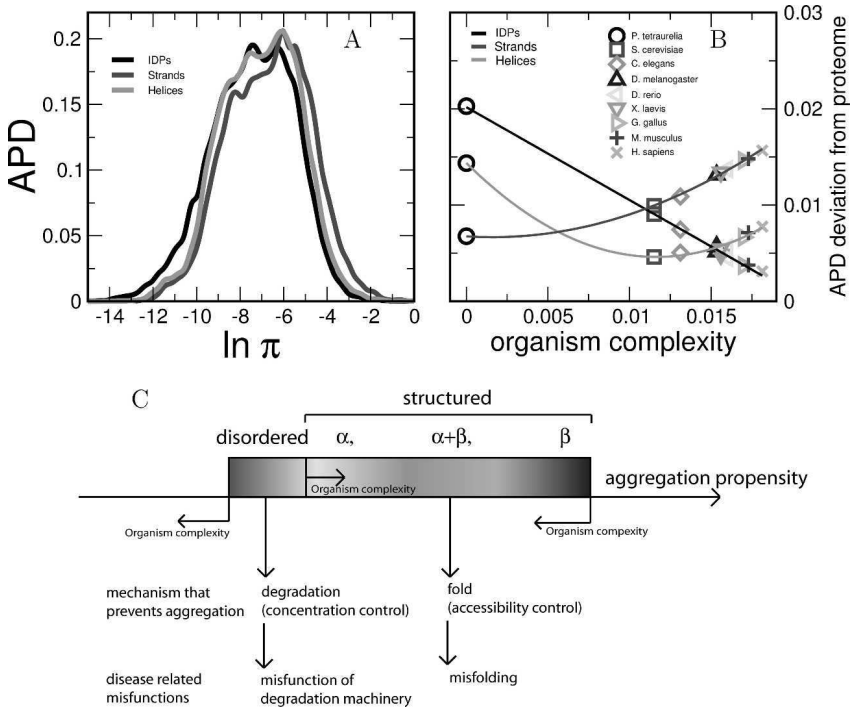

Figure 2. (*A*) APDs of five-residue stretches belonging to intrinsically disordered proteins (IDPs) or regular secondary structure elements within folded proteins and IDPs. (*B*) Deviation between the APD of entire proteomes and the APD of segments belonging to regular secondary structure or IDPs as a function of the organism complexity. The organism complexity is measured by the APD deviation from *P. tetraurelia*, d_{xP}. Solid lines are drawn solely to guide the eye. (*C*) From lower to higher eukaryotes, the decrease of β-aggregation propensity is related to the increase of intrinsically disordered proteins.

and IDPs are the least prone to aggregate, in agreement with Linding's analysis (Linding et al. 2004). Moreover, from lower to higher eukaryotes the APD deviation with respect to IDP decreases, while the APD deviation from strands increases (Fig. 2B,C). The APD deviation of helices does not follow a monotonic trend and slowly increases from *S. cerevisiae* to *H. sapiens*. Compared with strands, helices display a lower amount of aggregation stretches, but it has to be mentioned that the transition helix-strand generates amyloidogenesis in some proteins (Selkoe 1996; Prusiner 1997).

To quantify interspecies shifts of amino acid compositions in the APD tails, we fitted the amino acid frequencies as a linear function of the APD deviation from *P. tetraurelia*, d_{xP} (see Equation 4)

$$f_x^a = shift^a \; d_{xP} + cst^a \qquad (5)$$

where f_x^a is the frequency of the amino acid a in the proteome x, $shift^a$ is the slope of the fit, and cst^a is the intercept. The sign "+" or "−" of the $shift^a$ was interpreted as a measure for the depletion or the enrichment of the amino acid a from *P. tetraurelia* to *H. sapiens*. Shifts obtained from high-confidence fits (Pearson's correlation >0.80; Supplemental Material, Table 2) are

- Right tails, i.e., high propensity: Decrease of *Q*, *N*, *Y*, and *K* and increase of *L*, *V*, *A*, *W*, *R*, *H*, *G*, and *P*.
- Left tails, i.e., low propensity: Decrease of *K*, *I*, *F*, and *N* and increase of *P*, *A*, *G*, *R*, *S*, and *E*.

Interestingly, the decrease of *Q*, *N*, and *Y* in the right tails was already observed in higher eukaryote prion homologs of the yeast Sup35 prion protein (Balbirnie et al. 2001; Si et al. 2003a; Theis et al. 2003) and suggests that the trend does not affect only a specific family of proteins. In addition, we speculate that the increase of *L*, *V*, *A*, and *W* in the right tail is a consequence of the optimization of the "hydrophobic core" to stabilize the native state (Kellis et al. 1989; Richards and Lim 1993; Dill et al. 1995; Stefani and Dobson 2003).

The functional role of aggregation phenotypes in multicellular eukaryotes is still a matter of debate. Recently, it has been observed that the neuronal protein CPEB of *Aplysia californica* behaves like a prion switch that regulates long-term synaptic changes asso-

ciated with memory storage (Si et al. 2003a,b). The switch mechanism involves the aggregation of the CPEB N terminus, rich in *Q*- and *N*- repeats that are missing in mammalian isoforms of CPEB (Theis et al. 2003). Motivated by these observations, we analyzed the data set of proteins expressed in neurons (Supplemental Material , Table 1). For a given proteome, the neuronal APD perfectly overlaps with the APD of the total proteome (data not shown), indicating that neuronal proteins are a descriptive subset of the total proteome and do not follow any specific trend. We thus cannot draw conclusions on particular links between memory mechanisms and aggregation phenotypes.

It has been shown that the frequency of *N* and *Q* repeats does not represent an observable able to describe amyloidogenic trends of proteomes (Michelitsch and Weissman 2000; Osherovich and Weissman 2002). Our findings indicate that to quantify aggregation trends, it is crucial to use an observable, such as the β-aggregation propensity, which accounts for the aggregation contribution of all amino acids including positional information.

In conclusion, we have introduced a novel approach to compare proteomes, which is based on the statistical analysis of ordered-aggregation propensity. From *P. tetraurelia* to *H. sapiens*, we have shown that proteomes of higher and more long-lived eukaryotes contain fewer sequences with high β-aggregation propensity and are accrued in proteins with low β-aggregation propensity. We also observed that, compared with random proteomes, natural proteomes are enriched in proteins with low β-aggregation potential, as well as proteins with high β-aggregation potential. Such polarization is a consequence of the dual evolutive requirement of IDPs with low β-aggregation propensity, as well as proteins with a stable fold, which comes at the cost of higher β-aggregation propensity. In the future, we plan to use gene ontology annotations of proteins with high predicted β-aggregation propensity to obtain insights into the specific role of some of the amyloidogenic proteins of unknown function.

Electronic supplemental material

This section contains two tables: Table 1 contains information for databases used in the article (origin of data sets, number of entries of the databases, and number of stretches used in our analysis); Table 2 contains fitting parameters for the amino acid shifts (see Equation 5).

Acknowledgments

We thank Dr. A.G. Abebe and M. Cecchini for very interesting discussions. This work was supported by the Swiss National Science Foundation and the NCCR "Neural Plasticity and Repair."

References

Azriel, R. and Gazit, E. 2001. Analysis of the minimal amyloid-forming fragment of the islet amyloid polypeptide. *J. Biol. Chem.* **276:** 34156–34161.

Balbirnie, M., Grothe, R., and Eisenberg, D. 2001. An amyloid-forming peptide from the yeast prion Sup35 reveals a dehydrated β-sheet structure for amyloid. *Proc. Natl. Acad. Sci.* **98:** 2375–2380.

Bucciantini, M., Giannoni, E., Chiti, F., Baroni, F., Formigli, L., Zurdo, J., Taddei, N., Ramponi, G., Dobson, C.M., and Stefani, M. 2002. Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases. *Nature* **416:** 507–511.

Chiti, F., Calamai, M., Taddei, N., Stefani, M., Ramponi, G., and Dobson, C.M. 1999. Studies of the aggregation of mutant proteins in vitro provide insights into the genetics of amyloid diseases. *Proc. Natl. Acad. Sci.* **99:** 16419–16426.

Chiti, F., Stefani, M., Taddei, N., Ramponi, G., and Dobson, C.M. 2003. Rationalization of the effects of mutations on peptide and protein aggregation rates. *Nature* **424:** 805–808.

Dayhoff, M.O., Park, C.M., and McLaughlin, P.J. 1972. *Atlas of protein sequence and structure*. National Biomedical Research Foundation, Silver Spring, MD.

Dill, K.A., Yue, K., Fiebig, K.M., Yee, D.P., Thomas, P.D., and Chan, H.S. 1995. Principles of protein folding—A perspective from simple exact models. *Protein Sci.* **4:** 561–602.

Dobson, C.M. 1999. Protein misfolding, evolution and disease. *Trends Biochem. Sci.* **24:** 329–332.

Dunker, A.K., Brown, C.J., Lawson, J.D., Iakoucheva, L.M., and Obradovic, Z. 2002. Intrinsic disorder and protein function. *Biochemistry* **41:** 6574–6582.

Fitch, W.M. and Margoliash, E. 1967. Construction of phylogenetic tree. *Science* **155:** 279–284.

Gsponer, J., Habertuer, U., and Caflisch, A. 2003. The role of side-chain interactions in the early steps of aggregation: Molecular dynamics simulations of an amyloid-forming peptide from the yeast prion Sup35. *Proc. Natl. Acad. Sci.* **100:** 5154–5159.

Kellis, J.T., Nyberg, K., and Fersht, A.R. 1989. Energetics of complementary side-chain packing in a protein hydrophobic core. *Biochemistry* **28:** 4914–4922.

Kelly, J.W. 1998. The alternative conformations of amyloidogenic proteins and their multi-step assembly pathways. *Curr. Opin. Struct. Biol.* **8:** 101–106.

Koonin, E.V., Wolf, Y.I., and Karev, G.P. 2002. The structure of the protein universe and genome evolution. *Nature* **420:** 218–223.

Linding, R., Schymkowitz, J., Rousseau, J., Diella, F., and Serrano, L. 2004. A comparative study of the relationship between protein structure and β-aggregation in globular and intrinsically disordered proteins. *J. Mol. Biol.* **342:** 345–353.

Liu, J., Tau, H., and Rost, B. 2002. Loopy proteins appear conserved in evolution. *J. Mol. Biol.* **322:** 53–64.

Makin, O.S., Atkins, E., Sikorski, P., Johansson, J., and Serpell, L.C. 2005. Molecular basis for amyloid fibril formation and stability. *Proc. Natl. Acad. Sci.* **102:** 315–320.

Michelitsch, M.D. and Weissman, J.S. 2000. A census of glutamine/asparagine-rich regions: Implications for their conserved function and the prediction of novel prions. *Proc. Natl. Acad. Sci.* **97:** 11910–11915.

Oldfield, C.L., Cheng, Y., Cortese, M.S., Brown, C.J., Uversky, V.N., and Dunker, A.K. 2005. Comparing and combining predictors of mostly disordered proteins. *Biochemistry* **44:** 1989–2000.

Osherovich, L.Z. and Weissman, J.S. 2002. The utility of prions. *Dev. Cell* **2:** 143–151.

Osherovich, L.Z., Cox, B.S., Tuite, M.F., and Weissman, J.S. 2004. Dissection and design of yeast proteins. *PLoS Biol.* **2:** 442–451.

Prusiner, S.B. 1997. Prion diseases and the BSE crisis. *Science* **278:** 245–251.

Richards, F.M. and Lim, W. 1993. An analysis of packing in the protein folding problem. *Q. Rev. Biophys.* **26:** 423–498.

Rochet, J.C. and Lansbury Jr., P.T. 2000. Amyloid fibrillogenesis: Themes and variations. *Curr. Opin. Struct. Biol.* **10:** 60–68.

Selkoe, D.J. 1996. Amyloid β-protein and the genetics of Alzheimer's disease. *J. Biol. Chem.* **271:** 18295–18298.

Si, K., Giustetto, M., Etkin, A., Hsu, R., Janisiewicz, A.M., Miniaci, M.C., Kim, J.H., Zhu, H., and Kandel, E.R. 2003a. A neuronal isoform of CPEB regulates local protein synthesis and stabilizes synapse-specific long-term facilitation in aplysia. *Cell* **115:** 893–904.

Si, K., Linquist, S., and Kandel, E.R. 2003b. A neuronal isoform of the aplysia CPEB has prion-like properties. *Cell* **115:** 879–891.

Stefani, M. and Dobson, C.M. 2003. Protein aggregation and aggregate toxicity: New insights into protein folding, misfolding diseases and biological evolution. *J. Mol. Med.* **81:** 678–699.

Tartaglia, G.G., Cavalli, A., Pellarin, R., and Caflisch, A. 2004. The role of aromaticity, exposed surface, and dipole moment in determining protein aggregation rates. *Protein Sci.* **13:** 1939–1941.

———. 2005. Prediction of aggregation rate and aggregation-prone segments in polypeptide sequences. *Protein Sci.* (this issue).

Theis, M., Si, K., and Kandel, E.R. 2003. Two previously undescribed members of the mouse CPEB family of genes and their inducible expression in the principal cell layers of the hippocampus. *Proc. Natl. Acad. Sci* **100:** 9602–9607.

Ward, J.J., Sodhi, J.S., McGuffin, L.J., Buxton, B.F., and Jones, D.T. 2004. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.* **337:** 635–645.

Williams, R.M., Obradovic, Z., Mathura, V., Braun, W., Garner, E.C., Young, J., Takayama, S., Brown, C.J., and Dunker, A.K. 2001. The protein non-folding problem: Amino acid determinans of intrinsic order and disorder. *Pac. Symp. Biocomput.* **200:** 89–100.

Chapter 6

Conclusions

The mechanisms by which peptides and proteins form ordered aggregates are not well understood. In this study we focused on the physicochemical properties of amino acids that favour β-aggregation and suggested two parameter-free formulas to predict the aggregation rates. An essential element in the derivation of the models was the analysis of β-aggregating peptide sequences designed by genetic algorithm optimization in sequence space and molecular dynamics sampling of conformation space. The high correlations found between predicted and measured rates indicate that our models are able to describe *in vitro* experiments with high accuracy. Moreover, we were able to build the amyloid spectrum of a protein, identifying those segments which are involved in the β-aggregation. We found that mammalian and non-mammalian prion proteins show different amyloid-spectra, providing insights into the species barrier for the transmission of the prion disease. More specifically, we predicted a high amyloidogenic region corresponding to the segment `SNQNN` of the human prion which is absent in the chicken and turtle. At this stage, *in vitro* experiments represent the most important step to futher investigate the properties of amyloidogenic regions identified with our method.

Although in lower eukaryotes amyoloid fibrils could represent an inheritable phenotype related to specific cellular functions (as in the case of the neuronal protein CPEB of *Aplysia Californica*), the cytotoxicity of prefibrillar aggregates and their association with disease such as Alzheimer's, Parkinson's, Huntigton's, and prion disease, suggests that amyloid aggregates are generally dangerous for higher eukaryotes. We introduced a novel approach to compare proteomes using the statistical analysis of β-aggregation propensity. From *P. tetraurelia* to *H. sapiens*, we have shown that proteomes of higher and more long-lived eukaryotes contain fewer sequences

with high β-aggregation propensity and are accumulated in proteins with low β-aggregation propensity. We plan to use gene ontology annotations of proteins to obtain insights on the specific role of amyloidogenic proteins of unknown function. Disordered regions of proteins will be further investigated as determinant factors in preventing protein aggregation.

List of Figures

1.1 *Top plot*: Transmission electron microscopy of a mesh of amyloid fibrils assembled from human lysozyme negatively stained with uranyl acetate [59]; *Bottom plot*: Schematic drawing of the structural organisation of insulin fibrils. The image shows a fibril with four protofilaments wound around each other. In this model the core structure of each protofilament is a row of β-sheets, here running antiparallel [60]. 18

1.2 Calculated versus observed changes in aggregation rate upon mutation: AcP (28 triangles) and heterogeneous groups of peptide and protein systems including islet amyloid polypeptide, prion peptides, α-synuclein, amyloid β-peptide, tau, leucine-rich repeat and some model peptides (27 circles). 19

1.3 Amyloid protein precursor. The amyloid spectrum is averaged over a window of five aminoacids. The entire sequence is scanned by shifting the window by one residue at a time starting from the N-terminus. The analysis shows a major peak corresponding to the segment `LVFFA` at position 671. . . 20

1.4 *Inset:* Histogram of human polypeptide sequences as a function of β-aggregation propensity distribution at three different window sizes. *Main Plot:* Aggregation propensity distribution (APD) differences with respect to *H. sapiens* for complete proteomes of *M. musculus, D.melanogaster, C. elegans, S. cerevisiae* and *P. tetraurelia* (window size of 5 residues). Life-spans of organisms are reported in parentheses. 21

2.1 Aggregation of three heptapeptides: *a* parallel, *b* mixed, and *c* antiparallel conformation [1]. 29

2.2 Sketch of the genetic algorithm optimization (*LILA*). 29

2.3 Lila's performances for the parallel and the antiparallel β-sheet aggregation. For a total of 17 cycles, the fitness function is estimated to be the number of snapshots whose C_α-RMSD from the template is lower than 1 Å. 32

2.4 Trends of amino acid properties for the parallel and antiparallel optimizations. In the plots, the number of aliphatic, aromatic, charged, and polar residues is normalized by the length of the peptide and averaged over the population. . . . 33
2.5 Aggregation simulation of best parents. The variable Π indicates the number of snapshots whose C_α-RMSD from the parallel (black) or antiparallel (red) target structure is lower than 1 Å is used to build the histogram. 36
2.6 Aggregation of six replica-peptides `YFWLKFY`. Two variables are used to monitor the aggregation progress: The orientation parameter P_2 (defined in the range $[0, 1]$) and the number of C_α-contacts between peptides (defined in the range $[0, 35]$). . 37
2.7 Six replica peptides `YFWLKFY` forming a twisted fibril. 38
2.8 Parallel aggregation: The normalized fitness φ_p is used to train the β-aggregation propensity matrix P 39
2.9 Antiparallel aggregation: The normalized fitness φ_a is used to train the β-aggregation propensity matrix A. 40
2.10 Eigenvector analysis. The two eigenvectors v_a^{20} and v_p^{20} respectively correspond to the largest eigenvalues of the matrices A and P and indicate that charged residues stabilize the antiparallel configuration and aromatic residues stabilize the parallel configuration. 41

Acknowledgments

I thank Prof. Dr. Amedeo Caflisch for the stimulating conversations that allowed to reach my targets. I will never forget his support that helped me to improve my skills. A special thank to Dr. Andrea Cavalli who gratified me with his support and friendship during our collaboration. I cannot forget the synergetic interaction I had with Riccardo Pellarin, that I involved in all my projects. I would also like to thank Dr. Emanuele Paci, who supported me from my first day of work, and Enrico Guarnera, always present even if physically far. I am very grateful to Marco Cecchini for his advices and encouragements. I express my gratitude to Prof. Dr. Fabrizio Chiti and Dr. Michele Vendruscolo for providing experimental data and precious hints. I thank all the members of Caflisch's group: Francesco Rao, Dr. Michele Seeber Dr. Shaheen Ahmed, Raffaele Curcio, Fabian Dey, Danzhi Huang, Peter Kolb, Dr. Stjepan Jelakovic and Dr. Nicolas Majeux, Christian Bolliger and Dr. Alexander Godknecht, and Dr. Rainer Böckmann, Dr. Jörg Gsponer, Dr. Urs Haberthür, and Dr. Giovanni Settanni. Special thanks to Lindsey Munro, who helped me to correct my English.

I am grateful to all my friends and especially to Andrea Perucchi, Andrea Zampetti, Paola Sabatini Scalmati, Elena Aureli, Cristina Aureli, Fabian Rohner, Michael Zering, Joerg Domeisen, Claudio Cirelli, Francesca Albertini, Andrea Carminati, Vito Convertino, Fabio Krogg, Cinzia Finazzo, Fabio La Mattina, Luca Reggiani, Amsicora Giorgio Onnis, Giorgia Rama, Marco Miranda, Lapo Boschi, Aldo Ferrari, Erika, Nathalie, Misu, and Manu.

My studies have been possible for the patience, trust, esteem, and love of my wife Claudia. The encouragement of our families gave us the strength to live abroad, aware of their love.

Curriculum vitae

Personal Details

Surname: TARTAGLIA
Name: Gian Gaetano
Gender: Male
Date of birth: October 23, 1976
Place of birth: Rome, Italy
Nationality: Italian

Education

2001 - 2005	Employed as PhD student at the University of Zurich from November 2001 to September 2005 (Prof. A. Caflisch – Biochemistry Department University of Zürich)
1996 - 2000	Università la Sapienza (Rome, Italy) Master Degree ("Laurea") in Theorethical Physics with full marks (110/110)
	Theoretical thesis in Neural Networks: "Processi di Rinnovemento in Neurobiologia" *Renewal processes in Neurobiology* (Prof. B. Tirozzi – Dipartimento di Matematica – Università degli Studi di Roma)
1995	High School Liceo Classico "L.A. Seneca" – Rome (Italy) Diploma of high school with full marks (60/60)